Model-driven Study of Visual Memory

Final Report

AFOSR Award F49620-03-1-0376

July 1, 2003 – December 31, 2004

Robert Sekuler, Principal Investigator

Volen National Center for Complex Systems

Brandeis University, Waltham MA 02454

REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-05-

0276

1. REPORT DATE (MM/DD/YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
06/20/2005	Final Report	08/01/2003-12/31/2004

4. TITLE AND SUBTITLE
Model-driven Study of Visual Memory

5a. CONTRACT NUMBER

5b. GRANT NUMBER
F49620-03-1-0376

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Robert Sekuler

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Brandeis University

415 South St.
Waltham, MA 02454-9110

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

NL

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approve for Public Release: Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Short-term episodic visual recognition memory is crucial to success in many everyday activities. We synthesized concepts, insights, and methods from memory research, and from vision research, working within a coherent, quantitative framework for understanding episodic visual recognition memory. Seven experiments were carried out in two related sub-projects. One sub-project confirmed that high-dimensional stimuli (synthetic human faces) afford important insights into episodic recognition memory. The results were well accommodated by a summed similarity theory of recognition memory (Kahana & Sekuler, 2002). The second sub-project supported coordinated experiments on recognition memory and item identification (source memory). Source identification errors were deterministic rather than stochastic, and their causes were identified. Receiver operating characteristics (ROCs) compared recognition across experiments. Combining signal detection theory and a summed similarity model explained the unusual properties of the z-transformed ROCS.

15. SUBJECT TERMS
Memory, visual memory, computational model, human memory, faces, identity

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (include area code)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

7-13-05

Model-driven Study of Visual Memory

Final Report

AFOSR Award F49620-03-1-0376

July 1, 2003 – December 31, 2004

Robert Sekuler, Principal Investigator

Volen National Center for Complex Systems

Brandeis University, Waltham MA 02454

Contents

Overall objectives

Short-term, episodic visual recognition memory is crucial to the success of most activities of everyday life. The ability to recognize and evaluate recently-seen objects and events makes it possible to prepare and then execute appropriate actions in a timely fashion. Because this crucial capacity spans two distinctly different research traditions, vision and memory, it has received far less research effort than it deserves. With the interdependence of vision and memory firmly in mind, we are attempting to synthesize concepts, insights, and methods from memory research, and from vision research, working within a coherent, quantitative framework for understanding visual recognition memory, particularly for stimuli that resist rehearsal.

The reporting period covers the period of July 1, 2003 to December 31, 2004. During this project we made a strong progress on both of our research objectives, and on some supplementary objectives as well.

Objective One.

Examine episodic visual recognition memory for complex, high-dimensional stimuli, that is, synthetic human faces (Wilson, Loffler, & Wilkinson, 2002). Outside the laboratory, nearly all inputs to memory can only be defined in high-dimensional stimulus spaces. This is true of verbal material (words), natural sounds, and scenes. To date, recognition memory research supported by this project has focused on simple, low-dimensional stimuli, that is, 2- or 3-dimensional compound gratings. At issue is (1) whether NEMo (Kahana & Sekuler, 2002) and/or related models can account for recognition of more complex stimuli, which have higher dimensionality, (2) whether the general principles that govern memory for low-dimensional stimuli hold also for high-dimensional stimuli.

Objective Two

Examine relationship between recognition and identification (that is, source memory). Guided by our computational model, NEMo, and our prior findings, work in this objective developed and tested a new paradigm for studying the link between recognition of particular, previously seen stimuli, and the ability to identify the context in which the stimuli were encountered. Specifically, we collected and modeled concurrent recognition and identification judgments, applying signal detection and information theory analyses to the results.

Status of Effort

During the project period we collected substantial amounts of data and carried out model-driven analyses related to many of the project's objectives.

Objective One

Four experiments confirmed that high-dimensional stimuli (synthetic human faces) can provide useful insights into episodic recognition memory. Our initial data set replicated some, but not all, basic phenomena found previously with simpler stimuli. Using an eye-tracker funded primarily by this project, we examined the pattern of fixations that subjects made while they scrutinized and encoded the face stimuli for memory. We developed a reliable method for generating similarity judgments, which can be transformed via multidimensional scaling into a representation of perceptual distances among faces. A manuscript reporting this work is under review.

Objective Two

Three experiments examined connections between recognition and identification, a key attribute of of visual episodic memory. With compound gratings as study and probe items, subjects judged whether a probe had or had not been presented in the immediately preceding study series (a recognition judgment), and also identified the serial position of the study item that matched the probe (an identification judgment). Recognition and identification responses were expressed on a visual analogue scale. Approximately 75% of correct recognitions were accompanied by correct identification of the serial position of the study item that matched the probe. The suggests that recognition and identification are based on a common source of memory information. Misidentifications were attributable to two factors: perceptual similarity between the wrongly identified study item and the correct study item, and the temporal proximity of the wrongly identified item to the correct one. We used receiver operating characteristics (ROCs) to compare recognition performance across the three experiments. By combining signal detection theory (Wickens, 2002) and the summed similarity framework embodied in NEMo, we were able to account for the unusual slopes of our z-transformed ROCs.

Accomplishments/New Findings

Work on Objective One: Recognition Memory for Synthetic Faces.

Like other strongly-social creatures, *Homo sapiens* have a knack for recognizing previously-encountered conspecifics. Even a brief glimpse of some face can be enough to allow a viewer to recognize that the face had been seen before (Lehky, 2000). Going beyond this simple judgment of familiarity, additional, episodic information makes it possible to judge the circumstances under which that glimpsed face had been seen. To examine visual episodic memory for faces we adapted Sternberg's paradigm (1966, 1975), which can bridge visual psychophysics and memory, research domains that bracket visual recognition memory.

Our research is driven by a computational model that has successfully accounted for episodic memory with low-dimensional stimuli, compound sinusoidal gratings

that vary in spatial frequency and phase (Kahana & Sekuler, 2002). This model, NEMo (Noisy Exemplar Model), provides a framework for understanding how briefly-presented stimuli are represented in memory, and for identifying the way that stored memories are transformed into recognition judgments. Here, we apply NEMo to episodic memory for specially-designed human faces. Unlike the case for compound sinusoidal gratings, essential aspects of visual processing of human faces take place several synapses beyond the primary visual cortex. Because the primary visual cortex participates not only in visual encoding but also in visual memory and related phenomena (Magnussen & Greenlee, 1999; Kosslyn, Thompson, Kim, & Alpert, 1995; Klein, Paradis, Poline, Kosslyn, & Le Bihan, 2000), we were interested in the possibility that differences between visual processing of compound gratings and of human faces might produce corresponding differences in recognition memory for the two kinds of stimuli (Hole, 1996).

As readers may be unfamiliar with the class of face stimuli we used, and because the characteristics of these stimuli are central to our research, it is worthwhile to describe those stimuli in some detail. Wilson et al. (2002) devised a method for generating synthetic faces that are ideal stimuli for model-driven research on visual memory. Individual synthetic faces are derived from gray-scale face photographs by digitizing 37 key points: 14 points defining head shape, 9 points for the hairline, 4 points for eye locations, 4 points for nose length and width, 5 points defining the mouth and lips, and one point for brow height. Synthetic faces are then reconstructed from these 37 measurements and bandpass filtered with a 2.0 octave wide difference of Gaussians filter with a peak frequency of 10.0 cycles per face width (Wilson et al., 2002). Several studies have shown that such filtering is optimal for face recognition (Gold, Bennett, & Sekuler, 1999; Näsänen, 1999).

By design, synthetic faces eliminate textures such as skin, hair, wrinkles, etc, and focus instead on geometric characteristics of faces. However, this raises the important question whether synthetic faces are sufficiently accurate representations of the original faces to be useful in psychophysical experimentation. This question has been answered by requiring observers to identify the gray-scale photograph from which a synthetic face was derived in a four alternative forced choice experiment. The mean across five observers was 97.4% correct in matching between front view synthetic faces and photographs, and even for matching between 20 side view photographs and front view synthetic faces (or vice versa) performance averaged 90.7% correct (Wilson et al., 2002). As chance performance is 25% in these experiments, these data clearly demonstrate that synthetic faces capture salient aspects of individual face geometry. Furthermore, the data base captures known face gender differences: synthetic female faces have significantly smaller heads, rounder chins, thicker lips, and higher eyebrows than males. Finally, fMRI signals from the fusiform face area show that synthetic faces produce BOLD activation that is 85% as large as the original gray scale faces from which they are derived (manuscript in preparation). Using just 37 measurements strikes a balance between stimulus simplicity and encoding sufficient geometric

information to characterize individual faces. This essential individuality, which is important in episodic memory, is absent from some commonly used face stimuli, such as Brunswik faces (Brunswik & Reiter, 1937; Sigala, Gabbiani, & Logothetis, 2002; Peters, Gabbiani, & Koch, 2003).

The use of Wilson faces minimizes some problems associated with other collections of face stimuli (Duchaine & Weidenfeld, 2003). Furthermore, small graded differences among the faces allowed us to have a variability within individuals, and prevent subjects from learning and naming each face, which could subvert mnemonic reliance on visual information (for example, Ashby & Ell, 2001). To reinforce reliance on visual information *per se*, we limited rehearsal by giving subjects only a brief glimpse of each face, and then allowing only a short interval between successive faces.

To preview, our first two experiments establish key properties of visual short-term memory for synthetic faces. Experiment 1 characterizes the effects on recognition of serial position, list length, and similarity of the study items to probe. This experiment also explores memory decay resulting from the sequential occurrence of events, over and above the influence of similarities among items. Experiment 2 evaluates the effect of category-membership on recognition memory for synthetic faces, and evaluates the assumption that perceptual categorical effects are insignificant or absent. In Experiment 3 we employ a design that allowed us to use the NEMo model of visual short-term recognition memory to account for performance on individual lists (sets of stimulus items). The model was applied in several different alternative modes, for example expressing "similarity" either in terms of faces' physical coordinates, or in terms of perceptual coordinates, assessed using multdimensional scaling. Guided by NEMo, Experiments 3 and 4 allowed us to characterize the roles of (1) the similarity of probe to the list of study items, (2) the similarities among each of the list items, and (3) perceptual noise in predicting recognition judgments on individual lists.

Experiment 1

Our first experiment assessed the general suitability of multidimensional, synthetic face stimuli for use in studies of episodic recognition memory. We wanted to determine whether, when similar methods were used, findings from memory studies using low dimensional stimuli, such as compound gratings, could be replicated using higher-dimensional stimuli whose processing is known to engage regions of the brain that are not notably involved in processing compound gratings(for example, Kanwisher, McDermott, & Chun, 1997; Druzgal & D'Esposito, 2001, 2003)

We also empirically estimated the similarity-distance function for synthetic faces. This tuning function describes the relationship between recognition performance and pairwise Euclidean distance between faces. We planned to apply this empirically estimated similarity tuning function in modeling subsequent experiments; the aim was to reduce the number of free parameters in the model.

Stimuli The Wilson faces used in all our experiments were derived from photographs of three Caucasian females whom we designate A, B, and C. From $\mathbf{m}_A$, the vector of 37 measurements taken on actual face A, Wilson's procedure synthesizes a realistic version of that face in a stimulus space of high dimensionality ($n = 37$). Vectors of measurements taken on faces A, B and C were transformed so as to be mutually orthogonal, by Gram-Schmidt orthogonalization (Diamantaras & Kung, 1996; Principle, Euliano, & Lefebvre, 2000). Consequently, variation in one face's geometric properties is independent of the variation in geometric properties of the other two faces (Wilson et al., 2002).

After pre-processing and normalization, vectors of measurements from several different faces can be combined to generate $\mathbf{m}_{avg}$, the vector of measurements for an *average* face. To illustrate, the synthesized mean female face is shown at the left of Figure 1. This mean face is derived from a sample of 40 Caucasian female faces. Summing $k(\mathbf{m}_{avg})$ and $(1-k)(\mathbf{m}_A)$, for some k, $0 < k < 1$, generates a face that is a mixture of the mean face and some individual face, A. Allowing k to vary, $k = 0 \ldots 1$, generates a graded series of faces, which spans a continuum from the mean synthesized face (when $k = 1$) to a synthesized version of face A alone (when $k = 0$). The same operation also can generate a graded series of faces, which span the distance from the mean face toward any other face, here, toward B or C. The relatively small pairwise distances between faces, together with the orthogonalization of the face space, make it convenient to manipulate the similarity of one stimulus to another, a variable known to be important in recognition memory.

The graded series for faces A-C are shown in the upper three rows of Figure 1. Within each row, $(1-k)$ ranges from 0.04 to 0.20, in increments of 0.04, which means that each face differs from its nearest neighbor by approximately the mean discrimination threshold taken under viewing conditions similar to the ones we used (Wilson et al., 2002). In geometric terms, faces A-C lie along the mutually-perpendicular axes of a 3D space, with the mean face at the origin.

A final set of faces, D, was generated by averaging corresponding exemplars of A, B, and C. The resulting faces are shown in the bottom row of Figure 1. Geometrically, the faces in row D lie along the diagonal of face space, which means that the geometric properties of the faces in D are equally well correlated with the geometric properties of each of the other faces, A, B, and C. Figure 2A shows the geometric arrangement of all 21 synthetic faces in a space of three-orthogonal dimensions.

To prevent contamination of face recognition by emotional cues, our Wilson-faces incorporated standard generic shapes for features that would change shape as emotions are expressed, e.g., eyes, mouth and brows (Ekman & Friesen, 1975). In addition to having a consistent neutral expression, these faces were equated on dimensions such as contrast and mean luminance, which eliminates such attributes as aids to memory.

Procedure On each trial, one to four faces, the *study series*, were followed by a single *probe* face ($\mathbf{p}$). The faces in the study series comprised the items to be

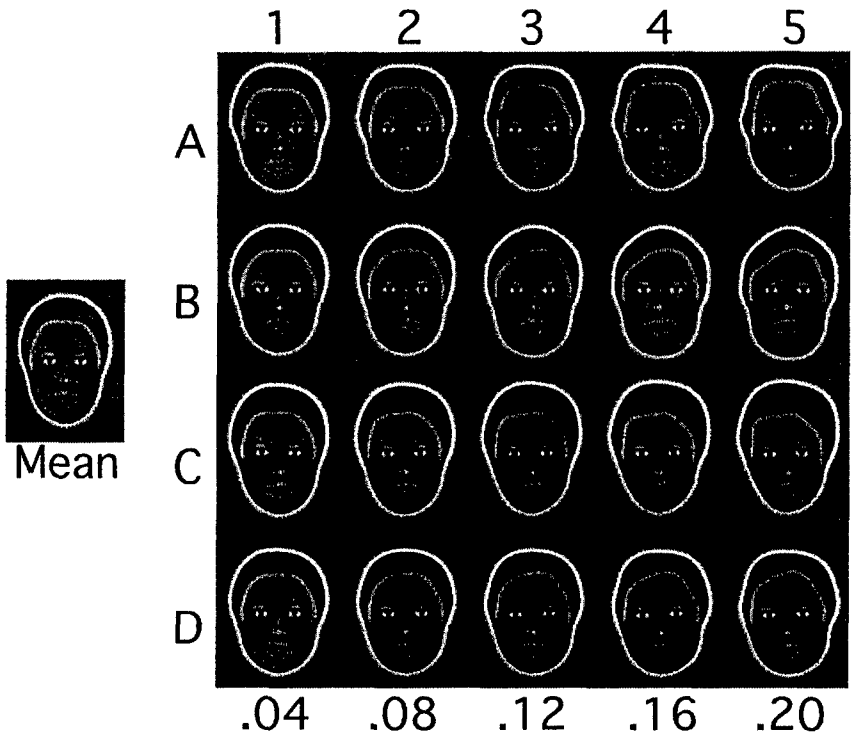

Figure 1. Face stimuli used in Experiment 1. Constructed after the method of Wilson et al. (2002), the stimulus labeled 'mean' is the average of 40 female faces. In the matrix of faces, rows A-C shows faces derived from three different faces; row D shows faces that are the means of faces A-C. Over the matrix columns $1, \ldots, 5$, faces deviate increasingly from the mean, by 0.04, in column 1, through 0.20, in column 5. For additional details, see the text.

remembered for that trial. Subjects judged whether **p** had been among the items in the study series. We use the term *target* to designate a **p** that had been in the study series, and the term *lure* to designate a **p** that had not been in the study series. Correspondingly, we can designate any trial as either a *target* trial or a *lure* trial. Because the study series varied from trial to trial, subjects were forced to base each "yes"-"no" recognition judgment on the items they had just seen.

Each study face was presented for 110 msec, with an inter-stimulus interval of 200 msec. The use of brief presentations was inspired by Wilson et al. (2002)'s use of this same duration in their studies of face discrimination, and by the fact that fairly detailed processing of a face can be completed within the first 100 msec of viewing Lehky (2000).

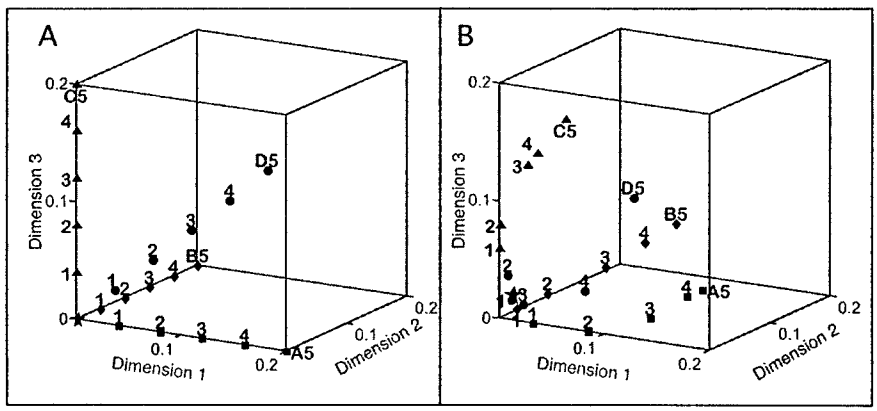

Figure 2. Representations of face stimuli in alternative three-dimensional spaces. In each panel, the star indicates the location of the mean face; squares, diamonds, circles, and triangles represent face A, B, C and D, respectively. The numbers 1-5 designate faces' distance from the mean; the numbers correspond to the columns in Figure . Panel A. The 21 face stimuli arranged according to the Euclidean distances between faces' physical descriptions, that is the physical distance of each face from the mean. Panel B. Arrangement of the stimulus faces in a three-dimensional perceptual space, using the MDS solution to position each face (Expt. 3). The MDS space shown here has been Procrustes transformed to bring its dimensions into line with those of the space shown in Panel A.

A warning tone followed the study series. Then, 1200 msec later, a **p** face was presented for 110 msec. To prevent potential within-category interference, in composing lists of study items only one study face could come from any one category of face, $A \ldots D$. For each study list, **p** was chosen at random from the entire set of faces, with two constraints. First, on half of all trials, **p** was forced to replicate one of the items in the study set (on half of all trials, **p** differed from all study items). Second, when **p** had been among the study items, with equal frequency it matched each of those items. Prior to the experiment a pool of 7477 unique stimulus series were randomly generated. Each subject was tested with a unique sample of series drawn randomly from this large pool of series, subject to constraints of list length and equal proportions of *test* and *lure* trials. Distinctive tones following each response gave subjects trial-wise knowledge of results.

Although there were small differences in size from face to face, each was approximately 5.5 degrees high by 3.8 degrees wide. To eliminate the usefulness of vernier-type cues, the vertical and horizontal position of each face was perturbed by adding a pair of random displacements drawn from a uniform distribution with mean = 12.6 minarc, and range = 2.1 - 23.1 minarc.

Subjects served in three, one-hour sessions of 576 trials each. During testing, a subject sat with head supported by a chin-and-forehead rest, viewing the computer display binocularly from a distance of 114 cm. Trials were self-initiated.

Subjects Subjects were five male and five female volunteers whose ages ranged from 19 to 30 years. They had normal or corrected-to-normal visual acuity as measured with Snellen targets, and normal contrast sensitivity as measured with Pelli-Robson charts (Pelli, Robson, & Wilkins, 1988). Subjects were naive with respect to the study's purposes.

Apparatus Stimuli were generated and displayed using Matlab and extensions from the Psychophysics and Video Toolboxes (Brainard, 1997; Pelli, 1997). Stimuli were presented on a 15-inch computer monitor with a refresh rate of 95 Hz, and a resolution set to 800 by 600 pixels. Routines from the Video Toolbox calibrated and linearized the display. Mean screen luminance was fixed at 36 cd/m^2.

Results Recognition memory performance is known to be influenced by the length of the study series and, on *target* trials, by the serial position of the study item that matches **p** (Sternberg, 1975; Murdock, 1982). We therefore examined performance as a joint function of these two variables. As study series lengthened, overall performance declined, $F(3, 27) = 49.57, p < 0.001$. The proportion correct responses did not differ significantly between the two types of trials, *targets* and *lures*, $F(1, 9) = 5.02, p > .05$.

For *target* trials, as list length varies, the effect of serial position is best understood by examining performance not as a function serial position *per se*, but as a function of *lag*, the number of study items between **p** and the study item that it matched (for example, Kahana & Sekuler, 2002). The data at the left side of Figure 3 show subjects' mean recognition performance as a function of lag for each study list length, LL=1 - 4. Note that lag=0 signifies that **p** matched the last study item in the series, that is, no study items intervened between the two. (Data at the right side of that figure represent performance on *lure* trials.) Proportion correct on *target* trials varies with lag position, with highest values achieved for lag=0, when the matching study item appeared at the very end of the series. With one or more study items intervening, performance declined (lag = 1, 2 or 3). This result constitutes a recency effect. Note that none of the series lengths showed a primacy effect, that is, an upswing in performance for the study item presented first. This serial position effect was consistent with the one reported by Kerr, Ward, and Avons (1999) using gray scale faces. We conjecture that the absence of any primacy effect resulted from the rapid presentation of stimuli, which inhibited rehearsal and therefore the appearance of a primacy effect (Ward, 2002).

Discussion For both *target* and *lure* trials, the proportion of correct responses declines as study series increase in length. This list length effect may reflect that as study lists lengthen, the recency effect becomes diluted, as more, less well-remembered, preceding items are averaged into the mix. The striking coincidence of the four curves at the left side of Figure 3 suggests that when *target* trials are equated for recency, there is little or no residual list length effect. This outcome is consistent with results

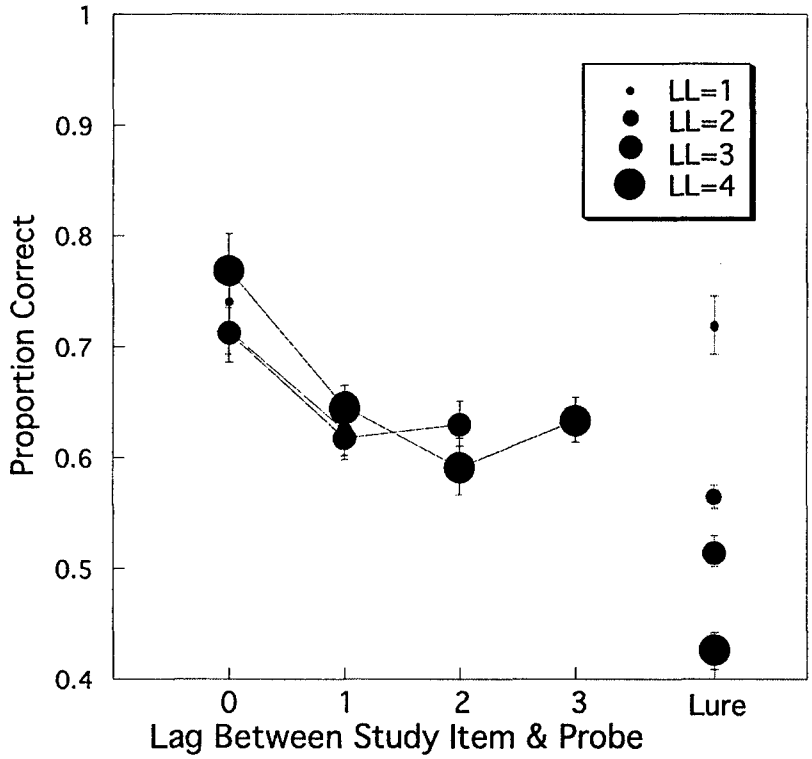

Figure 3. Recognition performance for *target* (left side of graph) and *lure* trials (right side). For *target* trials, proportion correct recognition is shown as a function of the lag between **p** and the study item that matched **p**. Curves are shown for study series 1, 2, 3 and 4 items long, with dot size corresponding to list length (larger dot for longer study lists). The four dots at the right side of the graph show proportion correct for *lure* trials. Error bars represent ± 1 standard error of the mean calculated according to method of Loftus and Masson (1994).

on recognition memory for series of 1-4 low-dimensional, compound gratings (Kahana & Sekuler, 2002).

Figure 3 shows that proportion of correct recognitions on *lure* trials fell with increasing numbers of faces in the study set. Before attributing this fall to some memory-related process, we must consider an alternative possibility. Because our study lists were generated at random, and because any face could appear just once in a study series, as series length grew, so too did the probability that one of the study faces on *lure* trials would be similar to **p**. If such similarity promoted false alarms (saying "yes" when no study item actually matched **p**), then that alone could generate an effect of series length, even in the absence of any degradation in memory with series length.

To examine this possibility, we reanalyzed those *lure* trials on which the study series had just one face. We reasoned that these one-item series would be least affected by any degradation in memory. These results were then fed into a model we describe as a Perfect Rememberer, which was then used to predict performance on longer study series, assuming that memory was perfect, e.g., unaffected by the number of faces in the study series. This simple model allows us to evaluate similarity's first-order effects on recognition memory performance. In addition, the Perfect Rememberer will set the stage for more detailed quantitative modeling, using NEMo a model to be presented later. NEMo will be introduced in conjunction with Experiments 3 and 4, which provided sufficient data to support a model-based analysis of individual lists of study and **p** items.

For treatment by the Perfect Rememberer we classified all one-item lure trials according to the distance between **p** and the study face. For this purpose, "distance" was taken as the Euclidean distance between a pair stimuli, i.e., the distances between faces in Figure 2A. We sorted the trials into bins according to the distance between the study face and **p**. Ten bins were generated with equal numbers of lure trials in each. All target trials were put into one bin for which the distance between **p** and the study face was zero. Finally, for each bin of trials we calculated the mean proportion of trials on which subjects recognized an item as a target. These means, which are plotted in Figure 4A, were used to predict what a hypothetical Perfect Rememberer would do when given series 2, 3 or 4 faces long.[1]

The Perfect Rememberer's expected responses are based on two simple assumptions. First, we assume that increasing the number of study items from the baseline of one has no effect whatever on memory, that is, there is perfect memory of each study face seen on a trial, with memory for any one study face being completely undisturbed by the presentation of other faces. Second, we assume that responses are determined solely by the one study face that is most similar to **p**, as defined in

[1]Because of the pre-probe delay in all study series, including series with just one face, we cannot say with confidence that the data for that shortest study series were entirely unaffected by decay of memory during that 1200 msec pre-probe delay. However, any such effect could not account for the changes in performance with the distance between **p** and study face, as shown in Figure 4A.

the space of 2A. With these assumptions, the proportion correct on lure trials can be obtained from the best-fitting curve shown in Figure 4A.

As this figure shows, the Perfect Rememberer is well described by an exponential function. This best-fitting exponential was

$$Pr("yes") = ae^{-\tau d(\mathbf{p},\mathbf{s})^c} \quad (1)$$

where (p, s) is the Euclidean distance between the probe and the stimulus face. The value of c was fixed at one, which produced a simple exponential function. The best fitting value of a, the y-intercept, was 0.84, and the value for τ, the rate at which the exponential decreased, was 9.2. These empirically-derived parameter values closely resembled the values estimated in (Kahana & Sekuler, 2002)'s application of NEMo to recognition results with simpler, compound grating stimuli. As the Perfect Rememberer depends upon the relationship between the physical distance and the corresponding behavioral outcome, this function comprises as a distance-similarity function. The parameters of the best-fitting exponential distance-similarity function for the Perfect Remember would be used later in model simulations of results from Experiment 3. The importation of this empirically-derived function was meant to reduce the number of free parameters in our modeling.

The gray bars in Figure 4B show the Perfect Rememberer's expected mean performance on lure trials for study series of length 2, 3 and 4. As expected, these values declined with series length, despite the fact they were came from a model that was memory-free. Note, however, that this memory-free decline differed substantially from the actual results of lure trials in Experiment 1. These latter values are shown by the black bars in the figure. Thus, we conclude that a substantial portion of the series length effect was caused, not by the similarity alone, but by some process that is memory-dependent.

The effect of series length obtained here with synthetic faces accords well with results from an analogous experiment on episodic recognition of lower-dimensional stimuli, compound sinusoidal gratings (Kahana & Sekuler, 2002). Specifically, with both kinds of stimuli, mean recognition performance declines with series length, there is no clear sign of retroactive interference effects, and a strong recency effect is seen, but no primacy effect.

Experiment 2

Experiment 1 demonstrated that one face's similarity to another influences episodic recognition memory. However, the precise nature similarity's role in memory remains to be clarified. When sets of study items were composed for Experiment 1, no more than one face was allowed to come from each category of face, A ... D. As a result, study items differed from one another in two ways: the categories from which they were drawn, and each face's distance from the mean face. Scrutiny of items that lie in the same row of Figure 1 suggests that for many faces, one can identify the

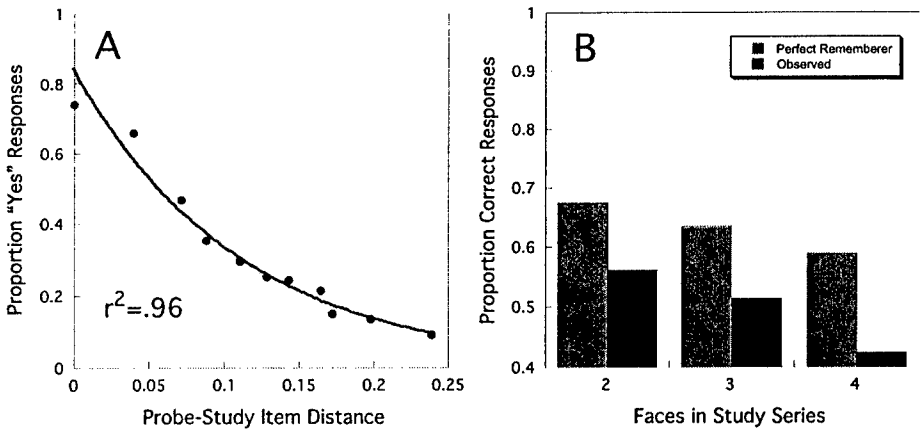

Figure 4. Panel A. Proportion correct recognition as a function of the difference between **p** and the study item in face space. Data are for study series of length 1. Error bars represent ±1 standard error of the mean of the Probe-Study item distance were smaller than the width of dots. Panel B. Predicted and obtained proportion correct recognition responses on lure trials for study series of length 2, 3 and 4. Predicted values assume perfect memory and judgments made on the basis of **p**'s proximity to that study item to which **p** is nearest, in face space.

characteristics each shares with others in its row. These shared characteristics permit the faces to be categorized.

Various studies have demonstrated that categorization can affect perception. When subjects learn to categorize stimuli, physical differences between categories gain importance relative to any within-category differences (for example, Goldstone, 1994, 1998; Levin & Beale, 2000). This is called the categorical perception effect, and is usually assessed under conditions in which subjects are allowed or encouraged to encode items together with the categorical attributes of those items. To minimize category effects in the present experiments, we kept stimulus duration and ISI brief. The model we intended to test later, NEMo, is mute about categorization, making no provision for such effects. If differences in category membership did affect recognition memory performance, NEMo would have to be significantly revised. Therefore, it was important to determine whether category membership played some role in our short-term visual memory paradigm.

In Experiment 2, we aimed to investigate the link between categorization and visual short-term memory. We compared memory for three-item long study series, which were generated according to two different rules. One rule forced all three study faces to come from different categories of faces, A ... D; the other rule allowed two or

even three study faces to come from the same category. We use the terms "between-category" and "within-category" to describe the series resulting from application of the two rules. Note that between-category series are formed according to the same rule that generated series in Experiment 1.

Methods

Apparatus and Stimuli The stimuli and apparatus were as in Experiment 1 except that multiple subjects were tested simultaneously, using computers in a classroom cluster. Although subjects did not use chin rests, they were encouraged to maintain a constant viewing distance of approximately 57 cm from their computer.

Subjects Twenty nine Brandeis undergraduates participated as part of a course requirement; during a session, after several trials of practice, each subject gave 436 trials. All subjects were naive to the experimental purpose, and none had taken part in Experiment 1.

Procedure The procedure was the same as Experiment 1 except that (a) study series always comprised three faces, and (b) study items and probes were presented for 250 msec each. During a testing session, between-category and within-category series were randomly intermixed, and occurred with equal frequency.

Results The gray bars in Figure 5A show the proportion correct on trials with within-category study series on target trials. Data are separated according to the serial position of the study face that matched **p**. The black bars in Figure 5A show results with between-category study series on target trials. Proportion correct responses were highest at the last probe position ($F(2, 56) = 31.6, p < .01$), replicating Experiment 1 (compare with Figure 3). A repeated measures ANOVA showed that the proportion of correct responses was higher for within-category series than for between-category series, $F(1, 28) = 16.5$, $p < .01$. This effect did not vary significantly with **p** position, $F(2, 56) = .054$, $p = .95$.

We suspected that the differences between the two conditions shown in Figure 5A might have overestimated the real differences between those conditions. On average, a study set whose members came from different face categories would tend to have members more different from one another, in Euclidean space, than would the average study set whose members all came from a single category of faces. Because recognition accuracy is influenced by stimulus similarity (Figure 4) we reanalyzed the data from Experiment 2, equating between- and within-category results for the similarity of study series members to one another.

For each stimulus series, we used the Euclidean distance between pairs of faces to calculate that series' mean pairwise distance among its study faces. We then ordered all stimulus series for each condition, between-category and within-category,

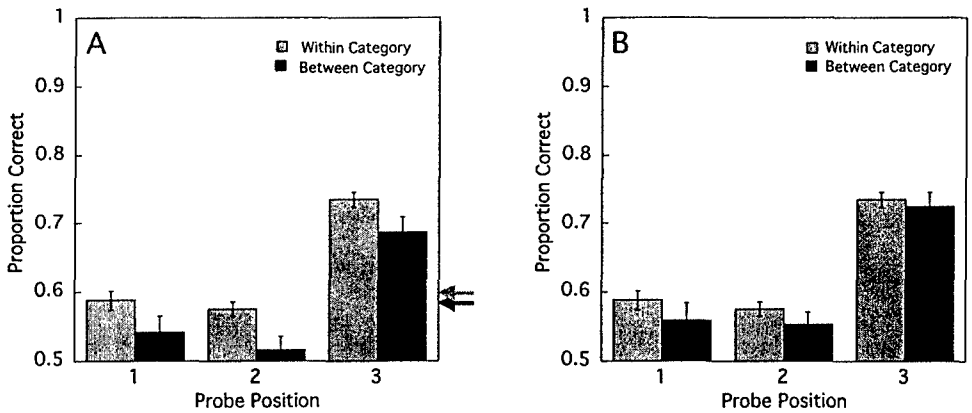

Figure 5. Panel A. Proportion correct recognition responses on target trials for probes matching study items in serial positions 1-3. Gray bars represent study series whose members were in the same category of face; black bars represent series whose members came from different categories. Arrows show proportion correct on lure trials. Panel B. Results after correction for *summed similarity.* In both panels, error bars represent ± 1 standard error of the mean, corrected for within subject variability (See Loftus and Masson (1994) for details

according to the series' mean pairwise distance. Next, from the between-category data we discarded all series whose mean pairwise distance exceeded the largest mean pairwise distance in the within-category series, and obtained equivalent mean pairwise distance for both conditions. The between-category trials that passed this test were used to recalculate the proportion correct responses shown for that condition in Figure 5B. Note this correction for similarity within a study series reduced the effect of category membership, and the difference between conditions were no longer statistically significant, $F(1, 28) = 2.42$, $p > 0.10$.

Discussion It appears that after the pairwise similarity of faces has been taken into account, the category to which the faces belonged exert a negligible effect on episodic recognition memory. Of course, it is impossible to deny that category membership *can be* consequential for various perceptual and memory tasks, under conditions different from our own, e.g., conditions of longer stimulus exposures, longer inter-stimulus intervals, or larger differences between faces. But our conditions clearly succeeded in minimizing effects of category membership. As a result of the null result, our subsequent experiments and modeling did not assign any special weight to category membership when similarity-related variables were being calculated.

Experiment 3

Experiment 1 suggested that the basic features of recognition memory for a series of briefly-presented synthetic faces may not be substantially different from memory for other, simpler stimuli tested under comparable conditions. For a deeper understanding of memory and other processes at work in this result, we exploited a summed similarity model that has successfully accounted for recognition memory with compound sinusoidal gratings. To maximize the power of the modeling we needed reliable empirical performance measures for each individual stimulus series that would be modeled. Such measures, in our memory paradigm, require ~30 replications per series for each subject (Kahana & Sekuler, 2002). To accommodate this many replications within a reasonable number of trials, we culled the stimulus series from Experiment 1, retaining only 60 series, each with three study faces, for use in Experiment 3. These 60 series spanned the full range of recognition performance.

Perceptual similarity among stimuli is central to the computational model we intended to apply to face memory. With compound gratings as stimuli, similarity can be defined by scaling stimuli in terms of subjects' difference thresholds for spatial frequency (Zhou, Kahana, & Sekuler, 2004). That same approach would be burdensome with face stimuli because the difference threshold varies substantially from one part of the space to another, with smallest difference thresholds in the neighborhood of the mean face (Wilson et al., 2002). Therefore, in addition to the native metric representations of synthetic faces, as represented in Figure 2A, we also used nonmetric multidimensional scaling (MDS) to characterize the perceptual space within which our face stimuli were located, and to quantify the distances between faces in that space. The data on which MDS was based came from oddity judgments made on simultaneously-presented trios of faces. To assess subjects' viewing behavior and strategy while viewing these faces, we followed up the main experiment by measuring the fixation behaviors of several subjects while they made oddity judgments like those made in the main experiment.

Subjects Two male and six female volunteers aged from 20 to 25 years participated in the main experiment, and two female volunteers aged from 20 to 22 years participated in the follow-up, eye tracking experiment. None took part in Experiment 1 or 2, and all were were naive to the purpose of this experiment. They had normal or corrected-to-normal visual acuity as measured with Snellen targets, and normal contrast sensitivity as measured with Pelli-Robson charts.

Procedure

Recognition Memory Of 60 different stimulus series, half were *target* trials, in which **p** replicated one of three study items; the remaining lists were *lure* trials, in which **p** replicated none of the study items. The timing of stimulus presentation was

identical with that of Experiment 1. Subjects participated in four one-hour sessions, 490 trials in each. The first 10 trials of each session were treated as practice and were eliminated from the data analysis; this retained 32 replications for each series and subject.

Multidimensional Scaling We used non-metric multidimensional scaling (MDS) to assess the perceptual similarity structure of the synthetic faces (Lee, Byatt, & Rhodes, 2000). The data required for such scaling were generated using the triangular method of trials (Ennis, Mullen, Frijters, & Tindall, 1989) On each trial, three faces were presented side by side, simultaneously for 500 ms, and subjects chose the one that seemed most different from the other two (Romney, Brewer, & Batchelder, 1993; Wexler & Romney, 1972). To minimize the possibility that vernier-type cues might contribute to the dissimilarity judgments, the vertical position of each was randomly offset by a sample from a uniform distribution spanning $\pm$ 16.8 min. Each possible stimulus pair ($Face_1$, $Face_2$) is presented with every other stimulus ($Face_3$). If stimulus $Face_3$ is selected as the stimulus most different from the others, then the remaining stimuli, $Face_1$ and $Face_2$, are deemed similar, either explicitly or by default. A similarity matrix is constructed by counting the number of times that a stimulus pair (e.g., $Face_1$, $Face_2$) is designated as "similar" when placed in combination with various other stimuli (e.g., $Face_3 \ldots Face_{21}$).

To control the number of trials required for the multidimensional scaling, we used a Balanced Incomplete Block design (Weller & Romney, 1988). For this design we generated triads of faces ("blocks"), whose members were drawn from the complete set of 21 faces. This selection was constrained so that each of the 210 *pairs* of faces occurred in the context of 30 triads. This arrangement meant that the 30 trials whose triads included any particular pair of faces were likely to have different faces as their third member. The displacement of three faces were randomly determined for each trial.

Subjects participated in three one-hour sessions, each with 710 trials. The first 10 trials of each session were treated as practice and were eliminated from our data analysis. The remaining 2100 triadic comparisons per subject were converted into dissimilarity matrices, which were processed by SPSS' ALSCAL and INDSCAL routines. All runs used an Euclidean distance model. A small, supplementary experiment was done to examine subjects' fixation patterns while they viewed a limited number of stimulus triads, and attempted to identify the face that was most different from the other two.

Apparatus and Stimuli In the main experiment, apparatus and synthetic face stimuli were the same as in Experiment 1. In a supplementary study of subjects' fixation behavior, stimuli were presented on a 19 inch monitor under the control of Matlab, and viewed with 57 cm viewing distance. The visual angles of presented faces were same as in Experiment 1. The monitor's refresh rate was 75 Hz; display mode

was set to 800 by 600 pixels. Eye position was recorded by an ASL Eye Tracking system (Model 504), a video based system that uses the pupil-corneal reflection to measure eye gaze location. Gaze was sampled at 60 Hz, and recorded to a spatial precision of just under 0.5 degrees visual angle. The apparatus monitored only the left eye. Prior to testing, eye position was calibrated using a standard, 17-target fixation pattern.

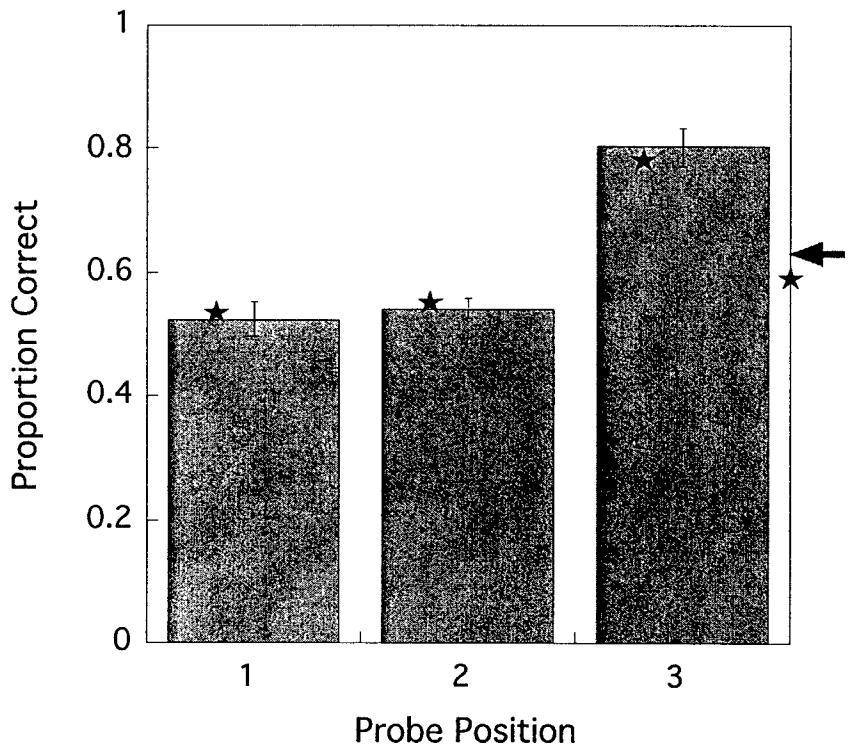

Figure 6. Proportion correct recognition responses on target trials for probe matching various study items. Error bars represent ±1standard error of the mean, calculated using the correction suggested by Loftus and Masson (1994). The arrow to the right of the graph represents the mean proportion correct on *lure* trials. Each star shows the proportion correct for the corresponding condition in Experiment 2.

Results

Recognition Memory Figure 6 shows the mean proportion correct as a function of **p**'s serial position. An ANOVA showed significant difference among those positions, $F(2, 14) = 24.43, p < 0.001$, and an *a priori* comparison revealed a significant recency effect, $F(1, 7) = 32.98$, $p < 0.01$. For comparison, corresponding results from Experiment 2 with the same series are shown as stars in the figure.

Multidimensional Scaling MDS solutions were obtained in spaces varying from 1-6 dimensions. r^2 values, which represent the proportion of variance accounted for in the scaled data, increased linearly as the number of dimensions varied from one to three; this increase saturated thereafter. The three-dimensional solution generated by MDS is shown in Figure 2B. Each symbol represents one synthetic face, and the pairwise distances between symbols represents the pairwise perceptual dissimilarity of the faces. Based on the values of r^2, we used three dimensional solutions in subsequent modeling. With the three dimensional space, Kruskal's Stress measure was 0.26, r^2 was 0.62.

Procrustes analysis (Dryden & Mardia, 1998) linearly transformed the matrix of values for the MDS solution to bring it into best conformity with the matrix of pairwise distances in the faces' physical space. The outcome, which is shown in Figure 2B, was based on the Euclidean similarity transformations of translation, reflection, orthogonal rotation, and isotropic scaling of points in the MDS solution. If the perceptual representation of these synthetic faces were identical to their physical representation, the post-Procrustes MDS solution would be perfectly congruent with the representation in Figure 2A, where faces A, B, and C are orthogonal to each other, and exemplars of face D lie on the diagonal.

Clearly, the Procrustes transformation does not eliminate all residual differences between the perceptual space, as represented by MDS, and the physical space. After the Procrustes transformation, the sum of squared residual discrepancies between the physical representations and the transformed-MDS representations was 0.26. We used Monte Carlo methods to cast this sum of squares into the same units that were used for the Euclidean physcial space (Figure 2A). Procrustes analyses were done on matrices of the faces' physical coordinates, which had been randomly perturbed to varying, known degrees, by the addition of independent, zero-mean, Gaussian random deviates to each of the three coordinates for each face. This operation was carried out 1,000 times for a number of Gaussian distributions with different standard deviations. From the mean residual sums of squares associated with each standard deviation value, we identified the random perturbation of face coordinates that produced the same residual sum of squares seen with the Procrustes transformation of the MDS solution. The standard deviation of the residual difference between the MDS solution and the original physical coordinates was equivalent to 6-7%, which corresponds to $\sim$1.5 times of the separation of neighboring faces within a single category of faces, A ... D.

To examine the residuals on a finer scale, the mean residuals between the MDS solution and the faces' physical coordinates were calculated and then sorted into bins according to the distance between faces in a study series and the mean of the 21 faces. These values are plotted in Figure 7. Note that the magnitude of the residuals grew with increasing distance from the mean face, suggesting that perceptual and physical representations of faces were most discrepant for the more extreme faces in our set.

As a further comparison between the MDS and physical representations of our 21

faces, we computed the vector angles between perceptual exemplars of A, B, and C. These vector angles are shown in Table 1. Faces just 4% away from the mean face were excluded from these calculations because in MDS space those faces clustered tightly around the mean face, which made angle measurements for those faces meaningless. The mean angles based on the 8, 12, and 16% data were roughly 90°, suggesting that the perceptual similarity space preserved much of the orthogonality that had been built into the physical representations of the faces. However, all of the angle estimates dropped appreciably when the 0.20 faces were included, which confirms the demonstration in Figure 7 that these extreme faces deviate most from Figure 2A's space.

Table 1: Angles Between Perceptual Representation of Faces A-C

Face Distance	AB Angle	AC Angle	BC Angle
8%	125.7	65.3	116.1
12%	83.2	74.2	80.7
16%	57.4	71.5	83.4
20%	46.5	54.9	63.3
Mean all distances	78.2	66.5	85.9
Mean 8% to 16%	88.8	70.3	93.4

Model NEMo, the model we applied to the recognition memory data, had previously been used to account for recognition memory with simple, low dimensional stimuli — compound gratings (Kahana & Sekuler, 2002). NEMo departs from the classic *summed similarity* models of item recognition (e.g., McKinley & Nosofsky, 1996; Nosofsky, 1986) by allowing recognition judgments to be determined not only by the similarity between the probe, $\mathbf{p}$, on one hand, and each study stimulus, on the other, but also by similarities among study items themselves. Given a series of L study items, $\mathbf{s}_1 ... \mathbf{s}_L$, and a probe, $\mathbf{p}$, NEMo responds "yes" if:

$$\underbrace{\sum_{i=1}^{L} \alpha\eta(\mathbf{p}, \mathbf{s}_i + \epsilon_i)}_{\text{Summed Probe-Item Similarity}} + \underbrace{\frac{2}{L(L-1)}\beta \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \eta(\mathbf{s}_i + \epsilon_i, \mathbf{s}_j + \epsilon_j)}_{\text{Mean Inter-stimulus Similarity}} > C_L \quad (2)$$

where $\eta(\mathbf{p}, \mathbf{s}_i)$ is the perceptual similarity between $\mathbf{p}$ and the i^{th} study item (see Equation 3, below); ϵ is a vector representing the noise associated with each stimulus dimension, α_i is the weight given the i^{th} study item, and C_L represents the optimal criterion for a series of L study items. To allow for the possibility that subjects' decision rule might incorporate inter-item similarity, NEMo adds together (i) *summed*

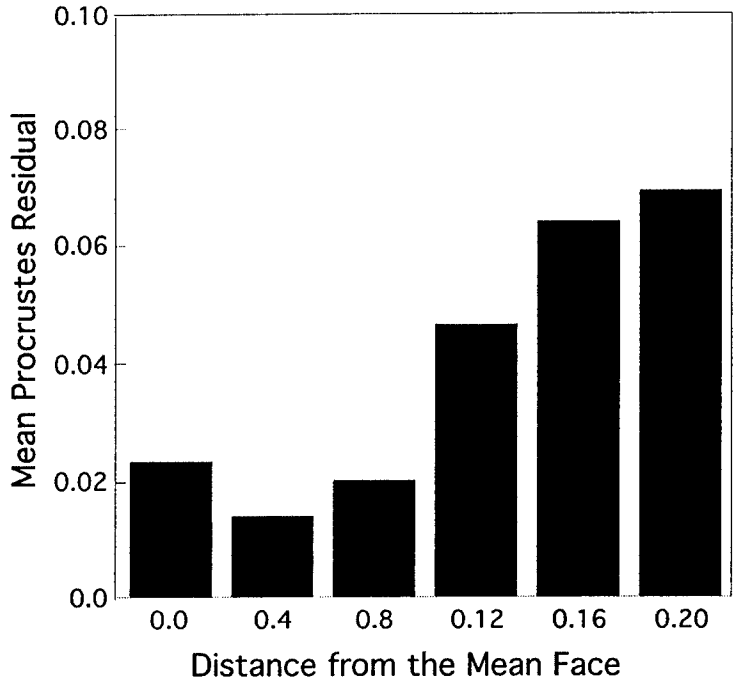

Figure 7. Mean distance between faces' representations in physical space and in perceptual (MDS) space, as a function of distance from the mean face.

similarity and (ii)inter-stimulus similarity, weighting the latter by a parameter β. If $\beta = 0$ the model reduces to a standard *summed similarity* model (Nosofsky, 1986) with noisy item representations (Ennis, 1988) and a deterministic decision rule. If $\beta < 0$, a given lure becomes more tempting when $\mathbf{s}_1 ... \mathbf{s}_L$ are widely separated from each other. Conversely, if $\beta > 0$, a lure becomes less tempting, and fewer "yes" responses are made, when $\mathbf{s}_1 \ldots \mathbf{s}_L$ are widely separated from each other.

In this model, the similarity, $\eta(\mathbf{s}_i, \mathbf{s}_j)$, between representations, $\mathbf{s}_i$ and $\mathbf{s}_j$, is given by:

$$\eta(\mathbf{s}_i, \mathbf{s}_j) = ae^{-\tau d(\mathbf{S}_i, \mathbf{S}_j)^c} \tag{3}$$

where d is the weighted distance between the two stimulus vectors, and τ, c and a jointly determine the form of the generalization gradient.

We defined similarity in both physical and perceptual (MDS) spaces, and than ran two parallel simulations of NEMo , one with each definition of similarity. Among parameters in Equation 3, we fixed $c = 1$, which implements a simple exponential generalization function. Previously, when c was a free parameter, its estimated values tended to be very close to 1 (Kahana & Sekuler, 2002). To reduce NEMo's free

parameters further, we used an independent, empirical estimate of τ and a. We estimated similarity in physical space from the data in Figure 4A, which we took as an empirical approximation to the similarity tuning function in physical space. After fitting an exponential function to the data in that figure, the function's exponent and y-intercept, 9.20 and 0.84, were used for τ and a respectively, in one set of model simulations. We then transformed the distances along the x-axis of Figure 4A using inter-face distances from MDS, and fit a second exponential function, this time to the transformed data. This second function represents the tuning function defined in perceptual space. The function's exponent and y-intercept, 11.14 and 0.91, were used as τ and a respectively, in a second set of model simulations. Finally, we fixed one other parameter, setting NEMo's criterion to 0.5, which is the rational, unbiased criterion that lies halfway between the means of summed similarity values generated on *target* and on *lure* trials. In studies of recognition memory for gratings (Kahana & Sekuler, 2002), subjects' mean criterion was found to be very close to this value.

Simulations and Application of Model We fit NEMo to subjects' recognition accuracy on each of the 60 different stimulus lists in Experiment 3. To find NEMo's best fitting parameters a genetic algorithm (Mitchell, 1996) minimized the root-mean squared-difference (RMSD) between observed and predicted recognition scores. In applying the genetic algorithm, a population of 1000 random parameter sets was allowed to evolve for 20 generations. Each parameter set was allowed to run for 1000 simulated trials for each stimulus list to produce an estimate of RMSD. Then each of the 500 least-fit parameter sets was replaced with a new parameter set at the end of every generation by randomly drawing each of their parameter values from one of the 500 best-fit parameter sets. The 500 best-fit parameter sets were mutated by a single, Gaussian parameter change with a standard deviation of 30% of a parameter's range.

We performed the entire model-fitting procedure twice. In the first application, inter-face similarity values, the value of τ, and a used in NEMo were defined by the physical distances between faces, that is, the differences among the parameter sets used to generate particular exemplar faces; in the second application, inter-face similarity values, τ, and a were defined by the MDS descriptions taken from subjects' *perceptual* space.

Results of Model Simulations: Fits to Mean Data We first fit *average* performance across eight subjects by using similarity in physical space and in *average* MDS perceptual space. Table 2 gives the best fitting model parameters for NEMo derived from the genetic algorithm. The column *Physical* shows the best parameters using stimulus distances in physical space, and the column *MDS* shows the best parameters using stimulus distances in MDS, perceptual space. The first three parameters, σ_1, σ_2, and σ_3, are the variances of noise distributions: one for each dimension of the three dimensional, perceptual space. σ_1, σ_2, and σ_3 correspond to

dimension 1, 2 and 3 in Figure 2. The next two parameters, α_1 and α_2, represent forgetting for the first and second item in a study series, respectively. (For the last item in a study series, α_3 was set to one.)

As explained earlier, β represents the importance of inter-item similarities. Its negative sign, $\beta = -0.53$ and -0.34 for the two simulations, indicates that when study items were similar to one another the model increased its tendency to treat a lure as a target. Note, finally, that the RMSD associated with the perception-based fit, 0.101, is somewhat smaller than the RMSD associated with the fit based on the faces' physical representation, 0.123.

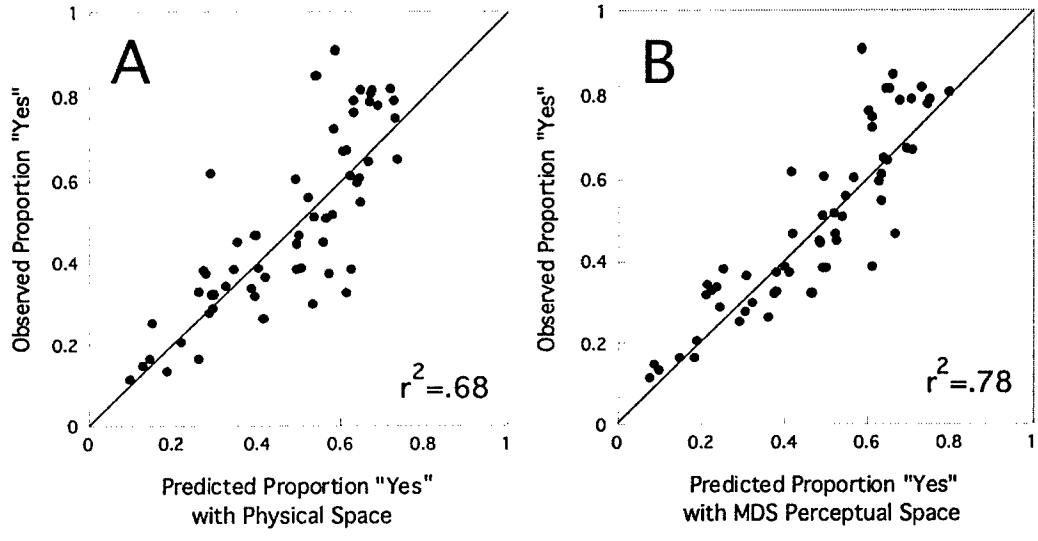

Figure 8. Proportion "yes" responses plotted against predictions from NEMo. Panel A. Interstimulus distances used in NEMo were taken from physical description of stimuli. Panel B. Interstimulus distances used in NEMo were taken from MDS solution.

Figure 8 shows the correlations between the predicted proportion of "yes" responses from NEMo and the observed mean proporion of "yes" responses. The predictions were made with similarity defined either by the physical representations (Figure 8A) or by the MDS solution averaged across all subjects (Figure 8B). With *average* data, NEMo produced a better account of the data when it incorporated perceptual similarity among faces. Physical and perceptual similarities produced $r^2 = 0.68$ and 0.78, respectively.

As just noted, NEMo's predictions tended to be more accurate when perceptual rather than just physical similarities were taken account of, but those predictions had a number of clear outliers. These were stimulus series on which the model failed badly, specifically series that deviated by 0.20 or more from the predicted proportion correct. To examine the cause of these failures, we scrutinized the makeup of these series. Of

the five outliers, we found that three contained two or more faces that deviated from the mean face by 0.20. To this outcome, Monte Carlo simulation assigned a probability $0.02< p <0.01$. So, faces with the greatest deformation relative to the mean face produced the largest errors in NEMo's predictions. As shown in Figure 7 and in Table 1, these extreme faces deviated appreciably from the orthogonal, physical representations of the faces. The perceptual transformations associated with the "strangeness" of these faces could be at fault. We suspect that "strangeness" essentially introduced an extra perceptual dimension specific to those stimuli, and because this extra dimension is limited to a small subset of faces, it would not be fully represented in the MDS solution.

Results of Model Simulations: Fits to Individual Subject Data Having fit the recognition results averaged over subjects, we fit individual performance by using physical representations of faces or by using individual MDS solutions for each of the eight subjects. One aim here was to obtain confidence limits for each parameter. Table 3 gives the mean and the standard deviations of best fitting parameters for each subject. Note that all values were consistent with those obtained by fitting the mean data. Of course, the RMSD values associated with individual subject fits were somewhat higher than the RMSD value associated with fitting the mean across subjects, because each subject's results would be based on relatively few trials per study series, which diminishes the reliability of the empirical data.

Parameters α_1 and α_2 were significantly smaller than α_3 for both simulations using physical space coordinates: $F(1,7) = 18.6$, $p < .01$ for α_1, $F(1,7) = 73.3$, $p < .01$ for α_2; with MDS perceptual space coordinates, the comparable values were $F(1,7) = 15.3$, $p < .01$ for α_1, and $F(1,7) = 61.0$, $p < .01$ for α_2. These inequalities, with α_1 and $\alpha_2 < \alpha_3$ represent forgetting. Additionally, a one-sample t-test showed that the mean β values were significantly less than zero for the model simulation with physical space coordinates, $t(7) = -3.51$, $p = .01$, and for the model simulation with MDS perceptual space coordinates, $t(7) = -4.15$, $p < .01$, which confirms *inter-item similarity*'s important contribution to recognition judgments.

To assess the difference between NEMo's predictions using physical descriptions and predictions using MDS descriptions of the faces, we compared two sets of simulations: one with physical descriptions, the other with MDS descriptions. The advantage of MDS perceptual description observed with NEMo's fit to *averaged* data disappeared. On the contrary, NEMo's predictions with physical descriptions were slightly better than NEMo's predictions with MDS descriptions, although the difference of RMSD values between two sets of simulations were not significant (matched-sample t-test, $t(7) = 2.13$, $p = .07$).

The reliability of MDS solutions Previously, we considered one explanation for the occasional failures in NEMo's fit: the perceptual distortion associated with the extreme items in our set of faces. Here, we take up another possible cause: within-

Table 2: Best fitting parameter values for NeMO's fit to the data

Parameter	Meaning	Physical	MDS
σ_1	Dimension_1 noise	0.033	0.032
σ_2	Dimension_2 noise	0.072	0.046
σ_3	Dimension_3 noise	0.045	0.046
α_1	Forgetting of 1^{st} item	0.47	0.57
α_2	Forgetting of 2^{nd} item	0.40	0.45
β	Interitem similarity	-0.53	-0.34
τ	Tuning function steepness	9.20	11.14
RMSD		0.123	0.101

Table 3: Means and SDs of best fitting parameters for individual subjects' simulations

Parameter	Meaning	Physical	*SD*	MDS	*SD*
σ_1	Dimension_1 noise	0.035	0.012	0.033	0.020
σ_2	Dimension_2 noise	0.053	0.032	0.061	0.022
σ_3	Dimension_3 noise	0.040	0.019	0.044	0.021
α_1	Forgetting of 1^{st} item	0.48	0.34	0.53	0.34
α_2	Forgetting of 2^{nd} item	0.37	0.21	0.42	0.21
β	Interitem similarity	-0.56	0.45	-0.67	0.46
τ	Tuning function steepness	9.20	-	11.14	-
RMSD		0.152	0.039	0.168	0.032

subject variability in the dissimilarity judgments on which the MDS solutions were based. To create a dissimilarity matrix for each subject, subjects observed three faces simultaneously for only 500 ms, and chose the face that appeared most different from the other two. The brief exposures were influenced by the stimulus duration in our recognition memory experiments. We reasoned that if subjects were to devote 110 msec to viewing each face in the trio, a total exposure of 500 msec would permit about two shifts in fixation, allowing all the faces to be viewed directly. However, the time pressures introduced by relatively brief stimulus presentations might also have contributed to variability in subjects' responses, which in turn would have introduced additional variability when those responses were transformed to dissimilarity matrices.

To assess the consistency of subjects' triadic judgments we computed two different mean MDS solutions. One solution was based on subjects' dissimilarity judgments on all odd-numbered trials (that is, first, third, fifth, etc.), the second solution was based on judgments from all even-numbered trials (that is, second, fourth, sixth, etc.).[2] For each three dimensional solution, the Euclidean distances between all face

[2]The balanced incomplete block design forced us to take this indirect approach. Differences in

pairs were taken, and the correlation calculated between pairwise distances from odd trials and pairwise distances from even trials. The results are shown as a scatterplot in Figure 9. Despite the brief stimulus duration, and despite the 50% reduction in the number of judgments on which each MDS solution was based, the pair of MDS solutions produced by this process had a relatively strong correlation with r^2=0.79, which supports the idea that triadic comparisons generate good, reliable measures of similarity.

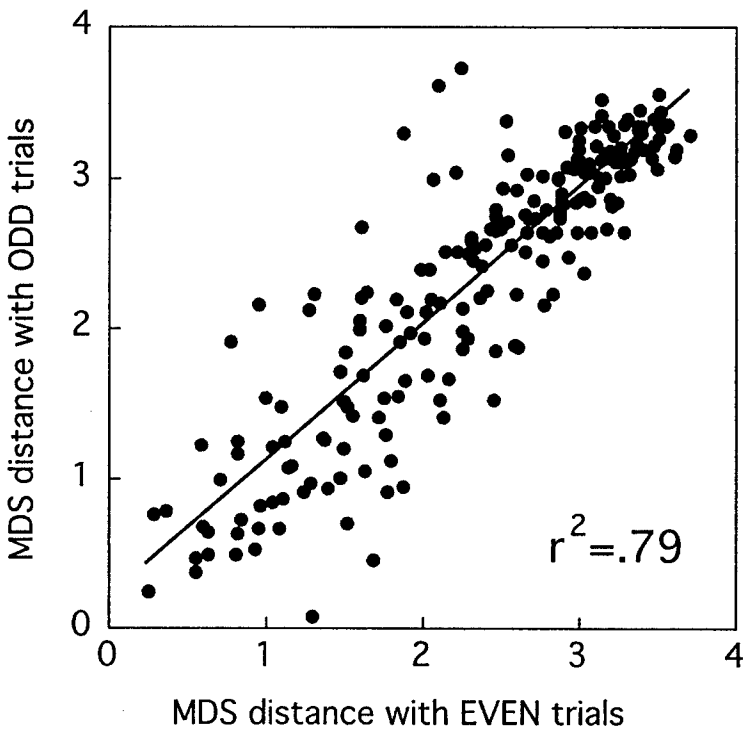

Figure 9. Distance between items in the MDS solution calculated by using even trials and them calculated by using odd trials.

Individual differences among subjects In addition to within-subject variability, between-subject variability, that is, various individual differences, could diminish the quality of model fits when NEMo was applied to data averaged over subjects. In the MDS solution, individual differences were represented by a vector of weights defined in an additional, individual difference space. Table 4 shows the

the makeup of triads on successive trials prevented us from comparing judgments themselves. As a result, our comparisons had to be mediated via MDS.

weights for each subject on each of the three dimensions, along with the r^2 values for each subject against the averaged MDS solution. The variation in weights and in r^2 values suggests that average MDS solution we obtained might not equally reflect different subjects' perceptual representations. For example, Subject 1 and Subject 5 differed from each other, both in their dimensional weights, as well as in their r^2 values. This variation in r^2 values might have come from differences in the consistency of individual subjects' similarity judgments. To test this hypothesis, we derived a within-subject measure of response consistency and rank ordered subjects' consistency measures and their r^2. A rank order test showed that the correlation between the two rankings was statistically significant (Kendall's τ= 0.84, p<.05), which was consistent with the idea that r^2 values reflected subjects' response consistency.

Table 4: Individual Subjects' INDSCAL Weights and r^2 Values

Subject	Dim 1	Dim 2	Dim 3	r^2
1	.49	.59	.41	.89
2	.45	.28	.38	.61
3	.54	.49	.41	.82
4	.63	.47	.34	.83
5	.15	.12	.13	.33
6	.45	.40	.47	.76
7	.58	.40	.32	.70
8	.41	.46	.51	.70

In a supplementary experiment, we measured several subjects' eye fixations while they performaed triadic comparisons. Of the 60 three-item study series used in main part of Experiment 3, we randomly chose 12 series for use here. Each trio of faces was presented five times in each of three spatial arrangements, e.g., xyz, yxz, or zxy. Repetitions of any trio were randomly interspersed over 180 total trials per subject. Three subjects participated in this supplementary experiment, one of whom had participated in the previous MDS experiment. On average, subjects made 1.93 fixation shifts per trial (SD = 0.09). On 47.7% of the trials, subjects seemed not to fixate every one of the three faces directly. However, for 88% out of all trials, subjects selected as most dissimilar one of the faces that they had directly fixated. As the distance between the center of one face and the center of a neighboring face was only about 5 degrees visual angle, subjects could easily use perifoveal vision (see Geisler & Chou, 1995) to identify the most different face and/or to decide the direction in which to shift gaze. Surprisingly, the consistency in subjects' judgments was not mirrored by consistency in the pattern of fixations elicited by a triad on successive appearances. In particular, the order in which faces were fixated seemed to be random, and not systematically related to the dissimilarity judgment that would be made.

Discussion

Multidimensional Scaling MDS requires as input a matrix of similarity or dissimilarity judgments. Researchers have taken various approaches to generate such matrices. For example, the input matrix has been generated by asking subjects to rate the distinctiveness of items presented one at a time (Valentine & Bruce, 1986; Valentine, 1991; Valentine & Endo, 1992; Lee et al., 2000), or to rate numerically the similarity of items presented in pairs (Nosofsky, 1991; Johnstone & Williams, 1997; Peters et al., 2003). We took a different approach, using triadic comparisons to produce the input matrix for MDS. We chose this method in part for its efficiency in generating many comparisons per pair of faces, and in part because the task might bear greater resemblance to our recognition memory task. For one thing, by encouraging subjects to distinguish among members of the briefly-presented triad, the task engaged the rapid and sometime subtle distinctions required in the recognition judgments. At the same time, variation in the triad's constituents from one presentation to another, mimicked the trialwise variation among study series.

Torgerson (1958) described other variants of the method of triadic comparisons, which require subjects to make several explicit pairwise judgments per trial. Note that the single explicit judgment required on each trial in our application actually implies that subjects have made one or more pairwise comparisons, although such comparisons are not made explicit. Letting the stimuli in the triad be i, j, and k, our subjects' identification of one item as most dissimilar could reflect evaluations of inequalities among $|i - j|$, $|i - k|$, and $|j - k|$. Of course, when subjects are not forced to make such evaluations explicit, one cannot rule out the possibility that, particularly with time pressures, subjects might make only some subset of all the pairwise comparisons. Although based on incomplete information, such judgments would be better and more consistent than random guesses.

Simulations with NEMo Episodic visual memory for synthetic faces was satisfactorily, though not perfectly, predicted by NEMo. That is, the results were consistent with the idea that each study item was stored as a noisy exemplar, and also with the idea that recognition decisions depended upon both *summed similarity* and *inter-item similarity*. When NEMo used MDS descriptions of face stimuli, the outcome was influenced by two sources of noise that were absent when NEMo used faces' physical descriptions. One source of noise came from any inconsistency in subjects' judgments of triads of faces. Although triadic comparisons were reasonably consistent, there is some obvious residual variability in those comparisons, as suggested by Figure 9. A second source of potential noise came from individual differences in similarity judgments. Table 4 shows each subject's weights on each of three MDS dimensions. It is clear that subjects differed in the weights they give to different dimensions. As a result, the mean MDS solution used in Figure 8 was not equally representative of every subject's own perceptual space. As these two noise sources were irrelevant to

the specification of faces in physical units, other things being equal, we might expect a larger error in NEMo's predictions when NEMo used face specifications from the MDS solution. This might explain why, when we simulated individual subjects' performance, we found no significant benefit of incorporating perceptual representations into the simulations.

Experiment 4

The simulations in Experiment 3 showed that visual memory performance could be predicted quite well using a model, which takes account of both *summed similarity* and *inter-item similarity*. Because the inter-item similarity term is a novel addition to the *summed similarity* framework, we sought a more direct demonstration that *inter-item similarity* actually was important in recognition. Therefore we designed stimulus series in which both *summed similarity* and *inter-item similarity* were varied. We expected that the responses produced by various combinations of the two factors would directly demonstrate the contribution of each factor, even without the mediation of a computational model.

Methods

Apparatus, Stimuli and Procedure The apparatus and stimuli were the same as in Experiment 2. The procedure was the same as Experiment 2 except that each trial's three study faces were forced to come from three different categories of faces, A...D. Study lists were first generated randomly. Because of the upper limit on possible pair-wise distances in physical space, distances among randomly selected faces tended to be small, which produced skewed distributions of summed similarity and inter-item similarity. Moreover, the randomly-generated lists, caused some covariance between summed similarity and inter-item similarity. To test wider ranges of both types of similarity independently, summed similarity and inter-item similarity had to be to distributed uniformly. To generate series that met this criterion, the distributions of two kinds of similarity were calculated after random generation of a set of series, and existing stimulus series were replaced by newly generated ones until we had a set of study series that satisfied the distribution requirement.

Subjects Twenty nine Brandeis undergraduates participated as part of a course requirement; during a single session, each subject produced 436 trials. All subjects were naive to the experimental purpose, and none had taken part in our other experiments.

Results and Discussion For each series on which **p** was a lure, we used the faces' physical coordinates to calculate *summed similarity* between **p** and all study items, and the *inter-item similarity*. For this purpose, similarity was defined by the

Euclidean distance in physical space, that is, the distance measures represented in Figure 2A. We sorted the trials into the cells of a 3×3 matrix, whose rows represented three levels of summed similarity, and whose columns represented three levels of inter-item similarity. At the end of the sorting, each cell of the 3x3 matrix contained 24 trials per subject. The proportion of "yes" responses was calculated separately for each of the nine cells in the matrix; these values are plotted in Figure 10. The parameter of the family of curves is inter-item similarity. The proportion of "yes" responses increased with *summed similarity*, but decreased with growth in *inter-item similarity*. A repeated measures ANOVA showed that both these effects were statistically significant, $F(2,56) = 85.6, p < .01$, and $F(2,56) = 7.08, p < .05$, for summed similarity and inter-item similarity, respectively. The interaction was not significant, $F(4,112) = 1.16, p = .33$.

Note that the directions of the two effects observed here reproduced the corresponding effects that we saw in simulations with NEMo. Of particular interest was the direct confirmation that inter-item similarity and summed similarity operate in opposed directions to influence recognition judgments, just as NEMo demonstrated they did.

General Discussion

Physical coordinates *vs* MDS solutions NEMo gave a good account of visual recognition memory performance for face stimuli. When average data across subjects were applied, the model's account was improved when face similarity was described in terms of the MDS solution, rather than in physical coordinates. However, the advantage of MDS solution was lost when NEMo was run on data from individual subjects. It is important to note that, with either perceptual or physical representations as input, NEMo had the same number of free parameters, therefore this null difference did not result from a difference in the number of free parameters.

One might think NEMo's predictions would be more accurate when model fits took account of perceptual representations. Earlier, we considered sources of variability in the triadic comparisons that were the basis for the MDS, perceptual representations. These sources of variability could have somewhat limited the quality of model fits. Another potential explanation for our failure to find improved fits with a perceptual representation is the lack of strong overall differences between the faces' physical descriptions and their perceptual descriptions. As shown in Table 1 and in Figure 7, the differences between two descriptions were small, except for those few faces that deviated most from the mean face. We speculate that had our set of faces been more heavily weighted toward extreme faces, the outcome might have differed.

We should note that the failure to gain an advantage from using a perceptual description of face stimuli is consistent with results from one other recent study. Peters et al. (2003) found that the performance of categorization models was either unchanged or even slightly diminished when face stimuli were described using MDS

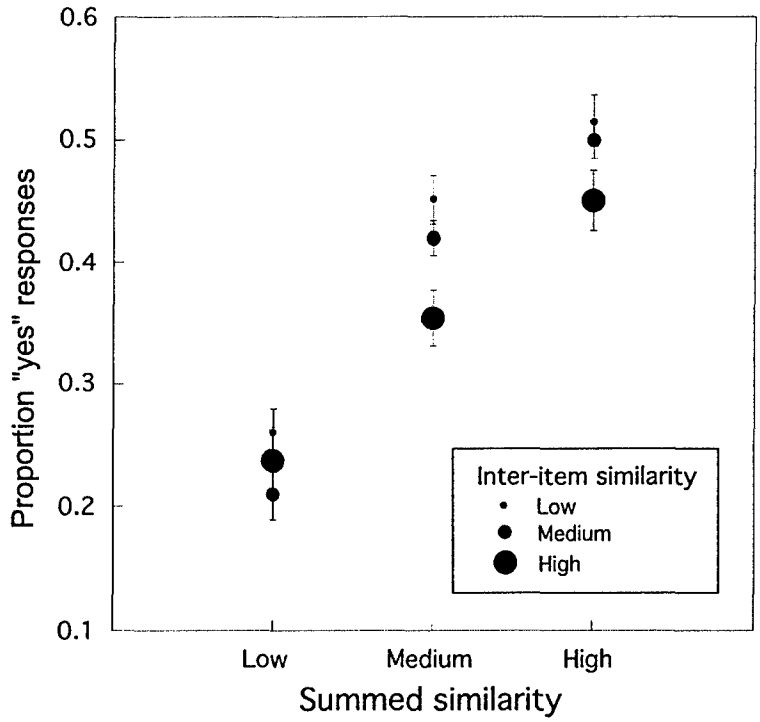

Figure 10. The proportion of "yes" responses as a function of *summed similarity.* Values are plotted separately for three different levels of *inter-item similarity.* The diameter of each filled circle signifies the magnitude of *inter-item similarity.* Error bars represent ±1 standard error of the mean, corrected for within subject variability (Loftus & Masson, 1994).

rather than a native, physical metric. In that study, subjects were trained to categorize schematic Brunswik-Reiter faces (1937), as well as slightly more elaborate, cartoon faces. These faces were defined in 4-dimensional physical spaces, which subjects learned to bifurcate using a linear, separable criterion. After learning the category membership of various exemplars, subjects' categorization was tested with mixtures of previously-seen and new faces. Peters et al. (2003) 's results suggested that subjects did not store the learned, long-term information as individual exemplars. We believe that this was quite different from the case in our experiments on episodic recognition, where subjects store individual exemplars, at least for the duration of a single trial.

Differences from compound gratings One of the purposes of this study was to investigate visual memory with higher dimensional stimuli, particularly by apply-

ing NEMo to memory for synthetic faces. The best fitting parameters obtained here preserved the general characteristics observed in Kahana and Sekuler (2002)'s studies with memory for compound gratings. For example, in both studies, α values captured the observed recency effects, and the significantly negative β values suggested that recognition judgments for both types of stimuli are modulated by *inter-item similarity.* Moreover, despite the fact that τ was set as a scale free parameter and a was set to one in the study with compound gratings, and τ were based on empirical estimates here, the obtained τ values were close between two studies (8.8 and 10.7 with compound gratings, 9.20 and 11.14 with synthetic faces).[3] It seems, then, that similar similarity-distance functions operate for both low-dimensional (gratings) and high-dimensional stimuli (synthetic faces).

However, memory for synthetic faces may differ in an important way from the memory for compound gratings. Even though we found recency effect with synthetic faces as well as with gratings, attributes of recency effect observed in this study differed from those found with gratings. With gratings, performance increased linearly across serial positions, from the beginning to the end of a study series (Kahana & Sekuler, 2002). Here, though, the serial positions preceding the last one produced essentially equivalent performance. This implies that with higher dimensional stimuli, instead of forgetting previously seen items gradually, the last seen item diminishes the memory for all previously seen items equally. This resembles a result reported previously (Phillips, 1974, 1983; Phillips & Christie, 1977).

Because procedures differed between the studies, we must be cautious in attributing various discrepancies in the results to differences in stimulus dimensionality. But we do believe that parallel studies of recognition memory for gratings, faces, and other high-dimensional stimuli, as well as comparable stimuli from other sensory modalities, will ultimately help us understand how stimulus dimensions contribute to human episodic recognition memory.

[3]We ran simulations with various τ values, and found that differences this small did not appreciably alter the model's resulting RMSD. So differences in the similarity-distance functions were not critical to the success of model fits.

Work on Objective Two: Coordinated Identification & Recognition

Many theories of memory distinguish between recognition that some item had been encountered previously (*item recognition*) and recognition of that item's previous context or source (*source recognition*) (for example, Johnson, Hashtroudi, & Lindsay, 1993; Hockley & Cristi, 1996). When multiple study items are presented successively, potential source information includes the items' temporal order (e.g. ?, ?, ?). Some researchers have suggested that item and source memory depend upon distinct neural substrates (Dobbins, Foley, Schacter, & Wagner, 2002; Rugg & Yonelinas, 2003). To assess the links between these two aspects of memory we generated coordinate assays of item and source memory using specially-designed stimulus materials and a novel paradigm.

Arguably, verbal study items, which are commonly used in memory studies, are less than ideal for theory testing. In particular, it is difficult to specify the representation of verbal materials in a metric space, a problem that can be significantly curtailed when using perceptually-defined stimuli (Kahana & Sekuler, 2002; Yotsumoto & Sekuler, in pres). For example, sinusoidal luminance gratings, which can be combined to synthesize more complex images such as textures and natural scenes, can be described in low-dimensional spaces whose principal axes are spatial frequency and orientation, dimensions that are extracted early in visual processing (Graham, 1989; Olzak & Thomas, 1999). Such stimuli, which can be designed to resist verbal encoding and rehearsal (Hwang-Grodzins, Jacobs, Danker, Sekuler, & Kahana, 2005), make it easy to manipulate similarity relations among stimuli, and to exploit those relations in computational modeling of memory (Kahana & Sekuler, 2002). The present work exploits the metric properties of memory stimuli in order to identify and quantify influences on source misidentification.

The present studies examine episodic memory for visual textures using both old/new recognition judgments (asking subjects to judge whether a probe was part of a just presented list) and source identification judgments (asking subjects to judge the serial position of the item within the list). Subjects expressed their judgments on an analog rating scale, which eased the generation of receiver operating characteristics (ROCs). ROC analysis, which is grounded in signal detection theory (Wickens, 2002), has proven fruitful in testing theories of recognition memory. Our own analysis builds on the fact that many models of recognition and source memory make explicit predictions about the shapes of ROCs (for example, Slotnick, Klein, Dodson, & Shimamura, 2000; Hilford, Glanzer, Kim, & DeCarlo, 2002; Rotello, Macmillan, & Reeder, 2004; Yonelinas & Levy, 2002).

Experiment 1

On each trial, subjects saw three briefly-presented study items. This series of study stimuli, whose members varied from trial to trial, was followed by a probe item (**p**), which either matched one of the preceding study items or differed from all

three. Subjects used an analogue rating scale to identify the serial position of the study stimulus that matched **p**; if no study item matched **p**, subjects registered that judgment with a *no* response. The analogue scale also allowed subjects to express their confidence in their judgments (Watson, Rilling, & Bourbon, 1964). These confidence judgments were used to generate receiver operating characteristics (ROCs), which were needed to test some theoretical predictions.

In Experiment 1, on 75% of all trials the probe item replicated one of the three study items. For terminological convenience, we use the term Target (T) to designate trials on which **p** replicated a study item, and the term Lure (L) to designate trials on which **p** did not replicate any of the study items. Additionally, we use t_i to designate the serial position in which the study item was replicated by **p**. Over all T trials, t_i equally often was the first, second or third item in the set of study items. Allowing T trials to occur 3x as often as L trials helped the examination of possible effects of t_i's serial position.

Stimuli Stimuli for each trial were drawn from a pool of compound sinusoidal gratings, each comprising superimposed vertical and horizontal sinusoidal luminance gratings. Each compound grating's luminance profile, $L_{x,y}$, was

$$L_{x,y} = L_{avg}\left[1 + \frac{A_1 cos2\pi f(x) + A_2 cos2\pi g(y)}{2}\right] \quad (4)$$

where L_{avg} is the mean luminance; f is the spatial frequency of the stimulus vertical component, (vertical frequency) in cycles per degree; g is the frequency of the horizontal component, (horizontal frequency); A_1 and A_2, the Michelson contrasts for the two components, were set to 0.2, a value well above detection threshold. To minimize edge effects, stimuli were windowed by a circular 2-D Gaussian with space constant of 1 degree visual angle. Before application of this window, a grating's width subtended 5 degrees visual angle.

Recognition performance for visual stimuli is influenced by visual as well as mnemonic factors, but we took steps to even out individual differences in vision. For each subject, we generated a unique pool of compound grating stimuli by crossing five vertical spatial frequencies with five horizontal spatial frequencies. Following Zhou et al. (2004), we measured each subject's visual discrimination threshold, and then scaled all of that subject's stimuli accordingly. This was meant to reduce the influence of individual differences in perception on measurements of memory performance. Frequency discrimination thresholds were measured using a staircase method in which subjects compared the spatial frequencies (2 cycles/degree) of two briefly presented gratings (750 msec duration each), which were separated by an inter-stimulus interval (ISI) of 400 msec. These same temporal conditions were used in our memory experiments. Frequency discrimination thresholds ranged from 4.3 to 13.1%, with a mean = 8.2, SD = 3.1. Vertical as well as horizontal spatial frequencies were allowed to assume values of 2 cycles/degree $\pm$ 3 or 6 times a subject's Weber fraction for changes

in spatial frequency. As the mean Weber fraction was 0.082, the spatial frequencies of the average stimuli were 1.51, 1.75, 2.0, 2.25, and 2.49 cycles/degree. Figure 11 illustrates the set of compound gratings that correspond to these typical values.

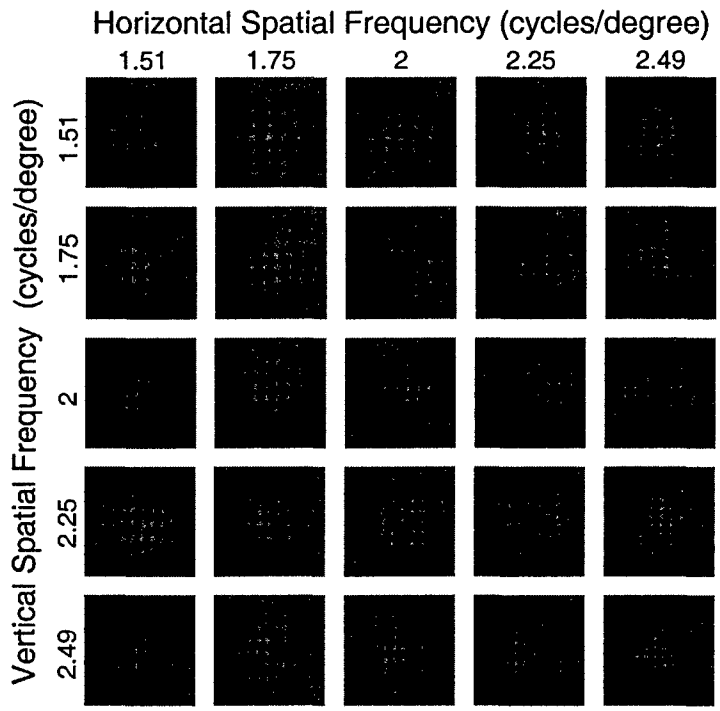

Figure 11. The average set of stimuli used in experiment. Within each row, vertical spatial frequency changes by three or six threshold units, decreasing and increasing from the mean of 2 cycles/deg, shown in the center of the stimulus matrix. Within each column, horizontal spatial frequency changes in the same way, again relative to the mean of 2.0 cycles degree.

Subjects Subjects were ten paid volunteers whose ages ranged from 19 to 28 years (mean=22.9, SD=3.3). Subjects' acuity, measured with Landolt C targets, ranged from 20/13–20/22, mean = 20/16.8, SD = 2.7; contrast sensitivity, measured with the Pelli-Robson charts (Pelli et al., 1988) ranged from 1.80–1.95, mean = 1.92, SD = 0.06.

Procedure Each trial's set of study items comprised three compound gratings, followed by a probe stimulus (**p**). Each of the three study stimuli ($\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$) was presented for 750 msec, separated by ISI's intervals of 400 msec each. Then, after a delay of 1000 msec, a warning tone sounded, and **p** was presented for 750 msec. One

second later, a response scale was presented, and remained visible until the subject's response had been registered. The response scale, shown at the left side of Figure 12, was used by subjects to report whether **p** had been in the study set, and, if so, which study item, first, second or third, was matched by **p**. This scale consisted of four selection arms, which were labeled "None," "First," "Second," or "Third." If **p** seemed to match one of the study items, subjects used the computer mouse to identify the arm that corresponded to the serial position (1st, 2nd or 3rd) of the study stimulus, $\mathbf{s}_1$, $\mathbf{s}_2$, or $\mathbf{s}_3$, which matched **p**. If **p** seemed to match none of the study items, the subject positioned the cursor on the arm labelled "None." In addition, subjects were encouraged to position the cursor in a way that expressed their confidence that they had selected the response correct arm. In particular, cursor positions near the intersection of the four arms signaled little confidence in the judgment; positions further away, toward to an arm's outer end signaled high confidence.

Once the subject was satisfied with the cursor's location, a click of the computer mouse button caused the computer to register the cursor's location along the response arm. No instructions were given about the speed with which subjects should respond. On average, once the scale was presented, a response was registered in about 2-3 seconds, which was sufficiently short that memory would not have decayed significantly (Kahana & Sekuler, 2002; Sekuler, Kahana, McLaughlin, Golomb, & Wingfield, 2005).

After each response, one of two tones sounded, providing feedback about response correctness. On *T* trials, feedback was contingent upon the response's identification component: feedback signalled whether the subject's response correctly identified which study item, $\mathbf{s}_1$, $\mathbf{s}_2$, or $\mathbf{s}_3$, matched **p**; like incorrect identification responses, a "None" response on a *T* trial brought feedback that the response was wrong. On *L* trials, feedback was contingent on whether the response correctly reflected that none of the study items matched **p**; all other responses, "First," "Second," or "Third," were followed by feedback that the response had been wrong.

Trialwise variation in stimulus spatial frequency forced subjects to base judgments on the most recently-seen study items; hence, the requisite memory can be described as episodic. Each subject was tested on 800 trials, distributed across four one-hour sessions.

The display's mean luminance was maintained at 17.8 cd/m^2, which prevented distracting luminance transients that would otherwise have accompanied change of stimulus. A subject viewed the stimulus display from a distance of 114 cm, head supported and steadied by a combination head rest and chin cup. Trials were self-paced. On each trial, $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ were sampled randomly without replacement from the pool of 25 stimuli had been were generated for that subject. On 75% of the trials, **p** replicated $\mathbf{s}_1$, $\mathbf{s}_2$, or $\mathbf{s}_3$, with equal frequency. On the remaining trials, **p** was chosen randomly from the 22 members of the stimulus pool that were not among that trial's three study items. These probabilities were explained to subjects prior to the experiment.

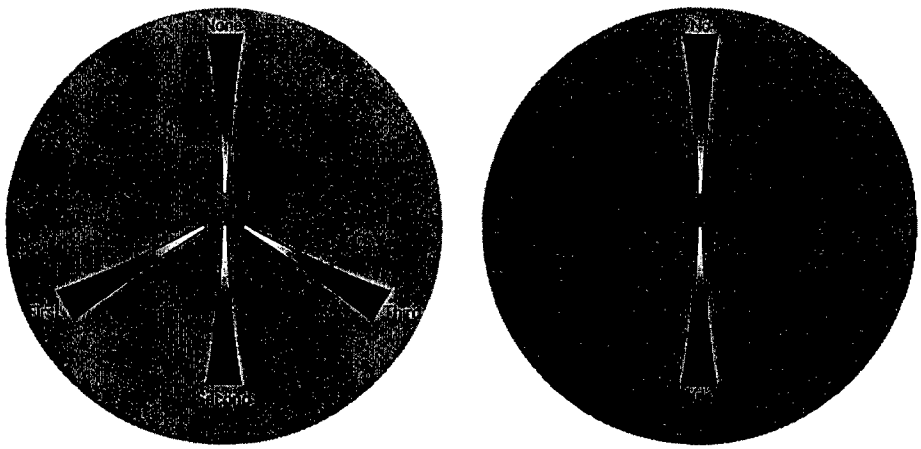

Figure 12. Left: Four-armed, analogue selection scale used by subects to express judgments in Experiments 1 and 2. Right: Two-armed, analogue selection scale used to express recognition judgments in Experiment 3.

Experiment 2

Although Experiment 1 examined source identification, its overall design was grounded in prior studies of recognition memory, in which no identification responses were taken (Kahana & Sekuler, 2002). Because those studies used balanced schedules of T and L trials, we were concerned that the unbalanced schedule used here in Experiment 1 might undermine comparisons between Experiment 1 and previous studies. Therefore, Experiment 2 replicated the conditions of Experiment 1, but with a balanced schedule of stimuli.

Subjects Five paid volunteers (aged 18 - 21 years, mean = 19.8, SD = 1.1) participated in this study. None had served in the preceding experiment. Subjects' acuity ranged from 20/13-20/20, mean = 20/16.6, SD = 2.7; contrast sensitivity ranged from 1.80-1.95, mean = 1.89, SD = 0.08; frequency discrimination thresholds ranged from 6.0-10.2%, mean = 9.2, SD = 3.1. All measurements used the same techniques as in the previous experiment.

Stimuli and Procedure This experiment used the same stimulus set as the previous experiment, with stimulus spatial frequencies again tailored to individual subjects' frequency discrimination thresholds. The proportion of T trials was decreased from 75% in Experiment 1 to 50% in Experiment 2. As before, with equal frequency $\mathbf{p}$ was made to match $\mathbf{s}_1$, $\mathbf{s}_2$, or $\mathbf{s}_3$. These probabilities were explained to the subjects prior to the experiment. There were no other differences between this experiment and its

predecessor. Each subject was tested on 800 trials, distributed across four one-hour sessions.

Experiment 3

In the preceding two experiments subjects made a source memory judgment on each trial, identifying the serial position of the study item that matched **p**, or responding "none". It was computationally simple to transform those source judgments into equivalent recognition memory, but we cannot ignore the possibility that the result might not truly correspond to recognition measured directly. How well does recognition measured indirectly, by means of transformed source judgments, correspond to recognition measured when no source judgment was required? To answer this question, for this experiment we modified the task used in the preceding two experiments, eliminating source judgments and requiring only recognition judgments.

Subjects Five paid volunteers (aged 18–20 years, mean = 18.6, SD = 0.9) participated. None had served in the preceding experiments. Subjects acuity ranged from 20/15–20/25, mean = 20/18, SD = 4.5; mean contrast sensitivity was 1.95; frequency discrimination thresholds ranged from 6.6–18.7%, mean = 10.6, SD = 4.7. Measurements were made using the same techniques as in the previous experiments.

Stimuli and Procedure The only difference between Experiment 2 and Experiment 3 was the judgment required of subjects, and the response selection screen on which judgments were registered. As shown in Figure 12 (see Right Panel), for Experiment 3, the oblique arms of the four-arm response display were eliminated, and subjects indicated only whether **p** had or had not been among the study items for that trial; no identification response was required. To remind subjects of the task, one arm of the response screen was labeled "yes," and the other arm was labeled "no." As before, participants signaled their confidence in each judgment by clicking on the appropriate arm, with distance from the center of the screen indicating increasing confidence. Subjects were informed that T and L trials would occur with equal frequency. In other respects, Experiment 3 was identical to Experiment 2. Each subject was tested on 800 trials, distributed across four one-hour sessions, and stimuli were tailored to individual subjects' frequency discrimination thresholds.

Overall performance measures

Experiment 1 For a basic level analysis, data were cast into a 4×4 stimulus-response confusion matrix. In this matrix, stimulus serial positions are designated 0 (for L trials), 1, 2 and 3, for each serial position; responses are described as "no" (**p** matched none of the study items), "yes-1" (**p** matched the study item in serial postion one), "yes-2," and "yes-3." Note that the proportions in the 4×4 confusion matrix, and the marginals, correspond to standard accuracy measures for recognition and source memory.

To assess various aspects of memory, values in the confusion matrix were combined in various ways. The first measure, the proportion of correct source identifications, is given by

$$P(Id) = P(r_i|t_i), where\ i \in \{1,2,3\}$$

where t_1, t_2, and t_3 represent the serial positions whose study items could match **p**; r_1, r_2 and r_3 are responses identifying **p** as matching the first, second or third study item, respectively.

The second measure, the proportion of correct recognitions, is given by

$$P(R) = P(r_i|t_j), where\ i,j \in \{1,2,3\}.$$

The third measure, the proportion of a correct source identification conditional upon a correct recognition, is given by

$$P(Id|R) = \{P(r_i|t_i)|P(r_i|t_j)\}, where\ i,j \in \{1,2,3\}.$$

Table 5 gives the the means and within-subject standard errors for these three measures. Also, the table's last row shows the proportion of all trials, T or L, on which subjects gave a source identification response, $\mathbf{r}_1$, $\mathbf{r}_2$, or $\mathbf{r}_3$ (on the remaining trials, subjects responded *no*). In standard recognition experiments, this value corresponds to the proportion of trials on which subjects would have responded "yes, **p** did replicate one of the study items."

Table 5: Summary Statistics for Experiments 1, 2, and 3 (Means and ±SeM)

Quantity	Expt. 1	Expt. 2	Expt. 3
P(Target)	0.75	0.50	0.50
P(Id)	0.68±.04	0.59±.04	-
P(R)	0.88±.02	0.78±.02	0.70±.03
P(Id|R)	0.77±.02	0.76±.03	-
P($\mathbf{r}_1 \cap \mathbf{r}_2 \cap \mathbf{r}_3$)	0.78±.02	0.57±.02	-

Note first that the P(R) is appreciably larger than P(Id), which seems to signify some loss of information in going from recognition (the sense that **p** had appeared in the study set) to source identification, which requires some memory of an study item's serial position. A rough estimate of this loss of information is found in the value of the table's fourth measure, P(Id|R), which falls somewhere between P(R) and P(Id). If there were no information loss, every correct recognition would be accompanied by a correct identification, producing P(Id|R)=1.0, rather than the value obtained, 0.77. Finally, the last row of Table 5 shows that on slightly more than three-quarters of all trials subjects responded that **p** matched one of the study items. Note that this value is a good match to the actual probability (0.75) which a **p** replicated a study item.

We take this as a sign that subjects used a near-optimal strategy, matching response probability to the actual probabilities of T and L trials.

For a more detailed examination of relationships among the trio of response measures in Table 5, the overall proportions were broken down according to the serial position of t_i, the serial position whose study item was replicated by **p**. Figure 13A shows the results. The proportion correct for L trials, 0.51±0.03 is shown by the single data point at the right side of the panel. The three serial position curves, one for each response measure, show a strong recency effect, with performance on T trials improving from t_1 through t_3. A repeated measures ANOVA confirmed this result, showing a significant main effect of serial position over all three curves in Figure 13A ($F(2,18) = 26.17, p < 0.01$). Moreover, the three curves departed significantly from parallelism, as confirmed by a significant interaction between serial position and type of performance measure ($F(4,36) = 5.25, p < 0.02$). Because values for P(R) are close to the upper limit of 1.0, an unambiguous of the interaction is not possible. interaction suggests that the information loss between recognition and identification, which was mentioned earlier, varies with serial position. Note that the P(Id) and PR) were calculated only from T trials, that is, only when **p** matched one of three study items. False alarm rates, the proportion of identifying L as T, were 0.12, 0.13, and 0.11 for s_1, s_2, and s_3, respectively. These false alarm rates did not differ reliably across serial positions ($F(2,18) = 1.60, p > 0.20$).

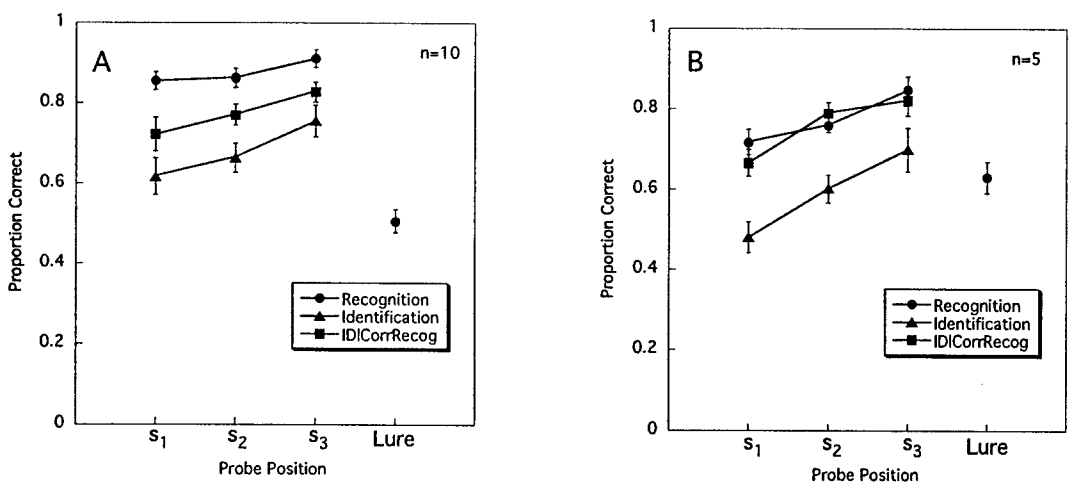

Figure 13. Proportion correct recognitions, identifications, and identifications given correct recogntions as a function of the serial position of the study item replicated by **p**. Shown also is the proportion correct rejection of lure stimuli (right side). Panel A: Results from Experiment 1; Panel B: Results from Experiment 2. Vertical bars around each data point represent ±1 within-subject standard error (e.g., Loftus & Masson, 1994).

Experiment 2 Table 5 shows the basic results for Experiment 2. Over all target serial positions, the proportion of correct recognitions was 0.78. The proportion of correct source identifications was just 0.59, while the conditional probabilty of a correct source identification given a correct recognition was 0.76, indicating essentially the same partial loss of source memory as in Experiment 1. Figure 13B shows the proportion of correct responses for L trials was 0.55 $\pm$.03. Additionally, as in Experiment 1, the serial position of the study item matched by **p** had a substantial effect ($F(2,8) = 20.70, p < 0.01$). False alarm rates were 0.09, 0.15, and 0.09 for t_1, t_2 and t_3, respectively.

Experiment 3 Recognition memory performance for Expeirment 3 was expressed as the proportion of correct recognition responses: 0.70$\pm$.026. As in the previous experiments, a repeated measures ANOVA demonstrated a main effect for the serial position of the study item that matched **p**, $F(2,8) = 11.62, p < 0.01$. We can compare the correct recognition measures from the second experiment, in which recognition was calculated by summing identification responses, to the forced-choice recognition measure in this experiment. A repeated measures ANOVA showed that the serial position results for Experiments 2 and 3 did not differ from one another, $F(2,16) = 1.74, p > 0.20$.

Receiver Operating Characteristic (ROC) analysis

To compare recognition measured by identification judgments (Experiments 1 and 2) and recognition measured directly (Experiment 3), we calculated the proportion of correct recognition responses produced by the two tasks. Recognition measures differed significantly between Experiments 1 and 2 ($p = .02$), most likely because of the experiments' different ratios of T to L trials. Given that T trials comprised 75% of all trials in Experiment 1, but just 50% of trials in Experiments 2 and 3, signal detection theory would predict these two values of P(R) to be ordered as they are. This hypothesis is bolstered by the observation that the overall proportion of both correct recognitions and false alarms were higher in Experiment 1 than in Experiment 2 (see Figure 13). Also consistent with this hypothesis is the similarity in P(R) for Experiments 2 and 3, 0.88$\pm$.02 and 0.78$\pm$.02 respectively, ($p = .18.02$).

The difference in stimulus schedule could have affected performance either by changing accuracy of memory, such as might come from differences in task difficulty and attentional demands, from a change in subjects' criterion, or from some combination of the two. To choose among these alternatives we generated receiver operating characteristic (ROC) curves from the judgments in each of the three experiments. In doing this, we exploited the confidence judgments provided by subjects' use of the continuous, analogue rating scale (Nachmias & Steinman, 1963).

Because our rating scale was analogue rather than comprising a fixed number of small categories, we sought an empirical estimate of how many useful categories – variations in confidence– were actually represented in subjects' use of the analogue

scale. After estimating that number, we used it to set the number of categories used to generate ROC curves from subjects' expressions of confidence. To determine the analogue rating scale's useable grain, we partitioned the rating scales into varying numbers of bins, and for each number we calculated the amount of information transmitted by responses (Garner & Hake, 1951; Watson et al., 1964).

Using the method of Watson et al. (1964), information transmitted was calculated for correct identication responses that had been sorted *post hoc* into varying numbers of bins. The number of *post hoc* response bins was varied from 2 to 12, with the constraint that for any subject, all bins contained equal numbers of responses. The information transmitted by these correct resonses grew with the number of response categories, but reached asymptote with no more than ten response categories.

Therefore, in generating ROC curves, we partitioned the analogue confidence responses into ten categories, with equal numbers of responses in each (see, Nachmias & Steinman, 1963). To test the similarity of results in Experiment 1 and Experiment 2, individual ROC curves were generated for each subject, and the area under each curve calculated using the trapezoidal rule for numerical integration (Wickens, 2002). In this process, separate curves were generated for $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$, and are shown in Figure 14. Figures 14A 14B show results from Experiment 1 and Experiment 2 respectively. $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ were plotted by filled circles, open squares, and filled triangles, respectively. The area beneath the curves was compared for each target serial position and for the two experiments. Although the main effect for target serial position was significant ($F(2, 26) = 47.32, p < 0.01$), neither the difference between experiments ($F(1, 13) = 0.63, p > .40$) nor the interaction of experiment and serial position were significant ($F(2, 26) = 1.87, p > .15$).

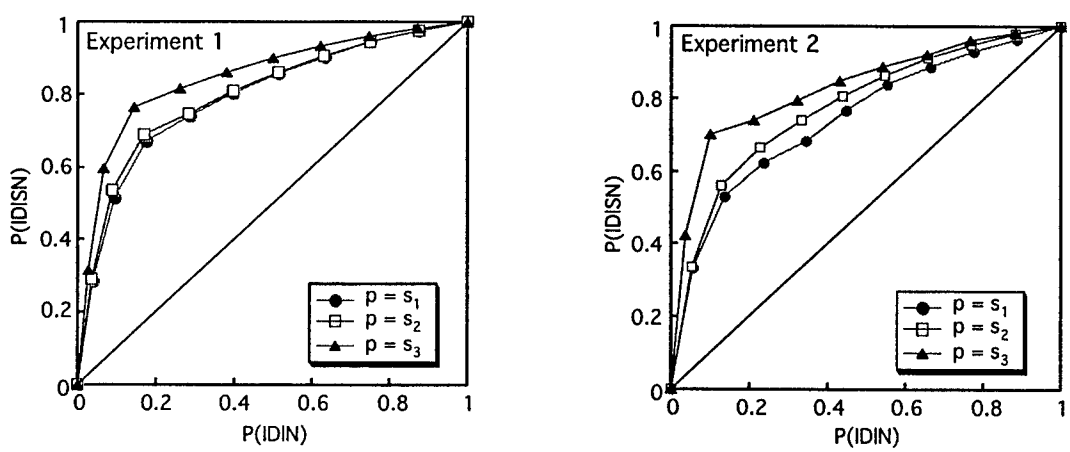

Figure 14. ROC curves based on source identification judgments for **p** in each of three serial positions. Panel A: ROCs generated from Experiment 1's results; Panel B: ROCs generated from Experiment 2's results.

Figure 15 shows the mean recognition ROC curves for all three experiments. In generating recognition ROC curves for Experiments 1 and 2, analogue confidence ratings for $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_3$ were aggregated. For the recognition ROC curve, derived from Experiment 3, we used the analogue confidence ratings associated with *yes* and *no* responses. These curves were generated for each subject, and the area under each subject's ROC curve was computed. The mean area under the ROC and the standard errors associated with that mean was 0.73±0.019, 0.75±0.027, and 0.68±0.037, for Experiments 1, 2 and 3, respectively.

A one-way ANOVA confirmed that areas under the three ROCs did not differ significantly from one another, ($F(2, 19) = 1.56, p > .20$). This outcome carries two important implications. First, the similarity of areas for ROCs from Experiment 1 and Experiment 2 ($p > .40$) suggests that differences in recognition performance between the two experiments (see Figure 13A, B) most likely arose from a change in criterion, rather than from a change in memory strength per se (Donaldson & Murdock, 1968). Second, the similarity of areas under all three curves supports the assumption that recognition measured directly, as in Experiment 3, is well approximated by recognition estimated by aggregating over the three separate identification responses, $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_3$.

Next, z-transformed ROC curves were generated for each subject by cumulating hit and false alarm rates and converting those cumulated values into standard scores (Figure 16). Figures 16A and B show the mean zROC curves for identification in Experiments 1 and 2, respectively; Figure 16C shows mean zROC curves for recognition in all three experiments. In a signal detection framework, the linearity of zROC curves suggests that the underlying noise and signal distributions are normal; if signal and noise were normally distributed, zROC curves would be linear in z-coordinates. A number of memory studies have reported linear zROC curves (Murdock, 1982; Donaldson & Murdock, 1968; Ratcliff, Sheu, & Gronlund, 1992). The slope of linear zROC curves reflects the relative variances of noise and signal distributions; if signal and noise distributions were normal and had equal variance, the resulting zROC's slope would be one. If the signal distribution had greater variance than the distribution on noise trials, zROC curves would have a slope less than one; the opposite relationship between variances would produce a zROC slope greater than one (Wickens, 2002)

The zROC curve generated for each experiment is well described by a linear function. Values of r^2's for the linear terms in a second order polynomial fit were > 0.95; addition of a quadratic term improved the fit by less than 0.02, which was not statistically reliable. The slopes for the zROC curves are shown in Figure 17. The slopes for zROC curves generated from source memory performance (Figure 17A) tended to be <1.0. Although several experiments have reported slopes of zROC curves for source memory ~1.0, those experiments differ in a number of critical respects from our own (Hilford et al., 2002; Glanzer, Hilford, & Kim, 2004). In contrast to these zROCs with slopes less than one, zROCs generated from recognition performance, either directly (as in Experiment 3) or indirectly (for Experiments 1 and 2) each had

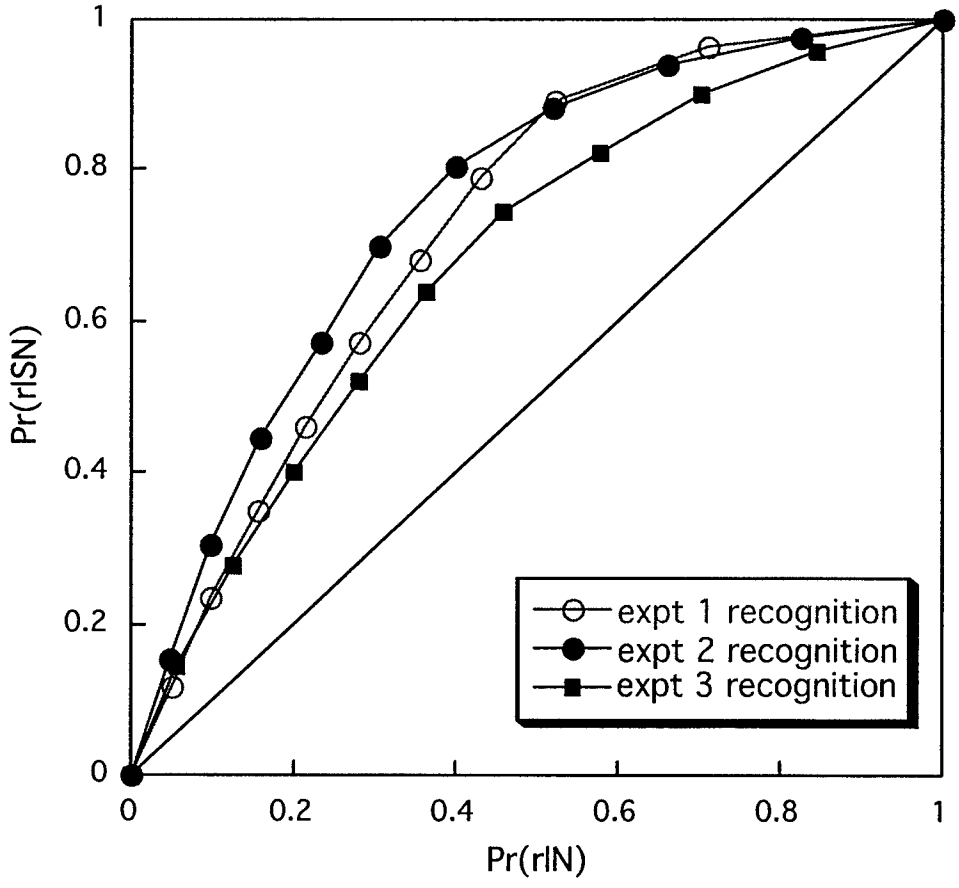

Figure 15. Receiver Operating Characteristics (ROCs) for recognition performance in Experiments 1-3. ROCs for Experiments 1, 2, and 3 are represented by open circles, closed circles, and filled squares, respectively.

a slope significantly greater than one (see Figure 17B); this was true even for Experiment 3, whose zROC had the lowest slope, $t(4) = 3.45, p < .03$). The slopes from Experiments 2 and 3 were not significantly different from one another ($p = .40$).

The few previously reported recognition zROCs whose slopes were greater than one were attributed to familiarity with the memory items (Ratcliff et al., 1992), or to differing levels of attention during encoding (DeCarlo, 2002, 2003). Although we cannot definitively rule out these possibilities, we believe that in our study a different influence produced zROC slopes greater than one. We take this up below, where we show that a summed similarity recognition model predicts this result.

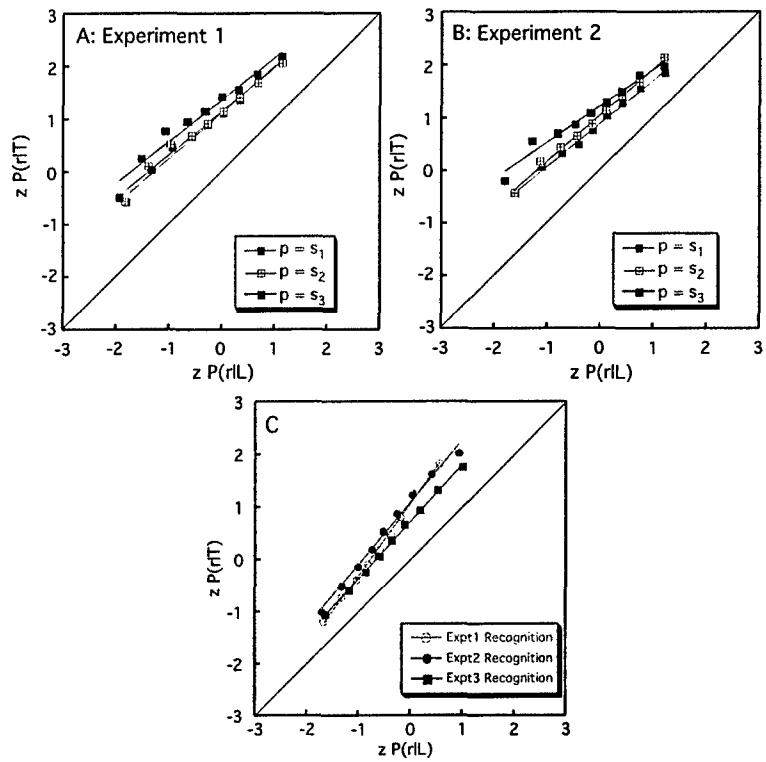

Figure 16. zROC curves generated from source memory and recognition performance. Panel A: zROC curves generated from source memory performance in Experiment 1. Panel B: zROC curves generated from source memory performance in Experiment 2. Panel C: zROC curves generated from recognition performance in Experiments 1, 2, and 3.

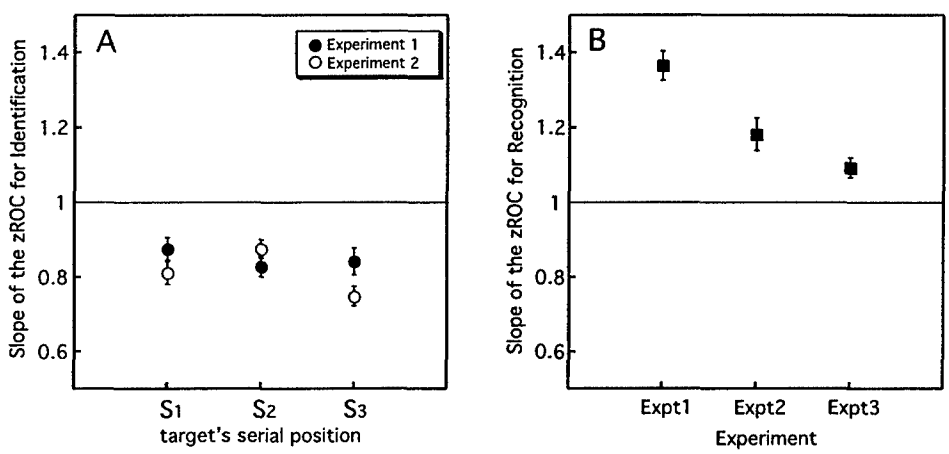

Figure 17. Panel A: zROC slopes for source memory in Experiments 1 and 2. Panel B: zROC slopes for recognition memory.

Stochastic vs. deterministic origins of source erors In both Experiments 1 and 2, subjects made many source misidentifications. On about 25% of all T trials subjects correctly rejected the *no* response, only to misidentify the serial position of the study item that had actually been replicated by **p**.

Errors in source judgments can be useful in defining the information that subjects use to make successful judgments. This is especially true when, as is the case here, the metric properties of stimuli afford the possibility of relating misidentifications to the characteristics of those stimuli. Misidentifications could have arisen in two distinct ways. Some or all of the misidentifications could have been entirely stochastic, reflecting random guesses made when subjects had no actual useable memory of what had been seen. Alternatively, misidentifications could have come from some deterministic process, for example, systematic errors associated with partial loss of serial position information (e.g., Lee & Estes, 1977).

Although pure, information-free guessing implies the existence of a threshold and is not consistent with non-threshold variants of signal detection theory, this hypothesis is useful as a starting point for analysis of misidentifications. In Experiment 1, 0.75 of all trials were T trials. As a result, information-free guesses on such trials would have been scored as correct recognitions 0.56 of the time, that is 0.75^2. Furthermore, because **p** was equally likely to match $\mathbf{s}_1$, $\mathbf{s}_2$ or $\mathbf{s}_3$, a random guess of serial position would have been scored as a correct identification on 0.1875 of guessed trials, that is, 0.5625/3. A source error, then, would have occurred, on twice this number of trials, 0.375 of trials on which the participant made an utterly random, memory-free guess that **p** matched one of the study items. Note that by definition, memory-free guesses cannot be systematically linked to any property of the study stimuli.

A misidentification could have come also, not from random guessing, but from one or more stimulus-related processes. With stimuli resembling those used in the present experiments, Kahana and Sekuler (2002) showed that visual encoding or study items was noisy, with each study being stored as a noisy exemplar. Suppose that a subject's judgments involved a process in which noisy (variable) exemplars of each study item were compared to **p**. This is the key computation in a large class of summed similarity models (Hintzman, 1988), including the one developed by Kahana and Sekuler (2002). By comparing the summed similarity, a scalar value, against, an appropriate criterion, the model generates"old-new" responses. However, this scalar value, alone, cannot generate identification responses. To account for such responses, we could augment the summed similarity framework by assuming that a judgment of serial position is determined by positional information of the study item whose memorial representation most closely matched **p**. Because of noise, on some trials, then, the memory for a study item that physically did not not match **p** might still more closely match **p** than does the memory for the study item that was a physical match for **p**. Additionally, a participant's memory for the matching study item might be sufficient to have produced a correct identification, but the participant misremembered the serial position in which that item had occurred.

To evaluate these two possible causes of misidentifications, we determined whether misidentifications might be explained by some spatial or temporal attribute of the stimuli. The spatial attribute was the pairwise spatial similarity between exemplars in the two-dimensional Euclidean space within which our stimuli were defined. The likely potency of this variable was suggested by Kahana and Sekuler (2002)'s account of false alarms in recognition judgments. For the temporal attribute, imagine that there were some orderly, non-random forgetting of serial position information. One form of non-random forgetting produces a "locality constraint," which promotes local rather than global errors. In the task at hand, items that occupied serially adjacent positions in a study sequence would be more likely to be confused with one another than would be positions that more widely separated in a sequence (Lee & Estes, 1977; Page & Norris, 1998). In the case at hand, with partial loss of serial position information, $\mathbf{s}_1$ would be more likely to be misremembered as $\mathbf{s}_2$ than as $\mathbf{s}_3$, and $\mathbf{s}_3$ would be more likely to be misremembered as $\mathbf{s}_2$ than as $\mathbf{s}_1$. This account is mute on errors arising from forgetting of $\mathbf{s}_2$'s serial position. Note that the duration of each study item (750 msec), together with the 400 msec separating successive items, should have been sufficient to minimize perceptual confusions between intervals, which suggests that misidentifications arise from failure of memory rather than failures of perception..

To compare competing stochastic and deterministic accounts of misidentifications, each participant's misidentifications were sorted into the cells of a 2×2 table. In constructing the table, we considered only trials on which $\mathbf{p}$ actually matched the first or last study item, $\mathbf{s}_1$ or $\mathbf{s}_3$; the remaining trials, on which $\mathbf{p}$ matched $\mathbf{s}_2$, do not lead to unequivocal predictions for misidentifications. The table's rows corresponded to two levels of a variable we call *spatial similarity*; the table's columns correspond to two levels of a variable we call *temporal similarity.* To generate the value of spatial similarity, we used a metric stimulus space in Figure 1 to calculate the Euclidean distance in spatial frequency between (1) $\mathbf{p}$ and the misidentified study item, and (2) $\mathbf{p}$ and the remaining study item that did not match $\mathbf{p}$. If the first of these two distances were the smaller, we categorized spatial similarity between $\mathbf{p}$ and misidentified item as "high;" otherwise, we categorized spatial similarity as "low." For temporal similarity, we categorized misidentifications according to whether the error in identification represented a shift of either one (high similarity) or two (low similarity) serial positions. For example, if $\mathbf{s}_2$ were misidentified as matching $\mathbf{p}$, when the actual matching study item was $\mathbf{s}_3$, this error of one serial position was categorized as high temporal similarity; if $\mathbf{s}_1$ were misidentified as matching $\mathbf{p}$, when the actual matching study item was $\mathbf{s}_3$, the error of two serial positions was categorized as low temporal similarity. A factorial cross of spatial and temporal variables produced four combinations of spatiotemporal differences between $\mathbf{p}$ and the misidentified study item. Trials involving a match to $\mathbf{s}_2$ were omitted from this analysis, because clear predictions involving those trials could not be made within the theoretical framework used here.

Figure 18 shows the proportions of source errors falling into each of the four

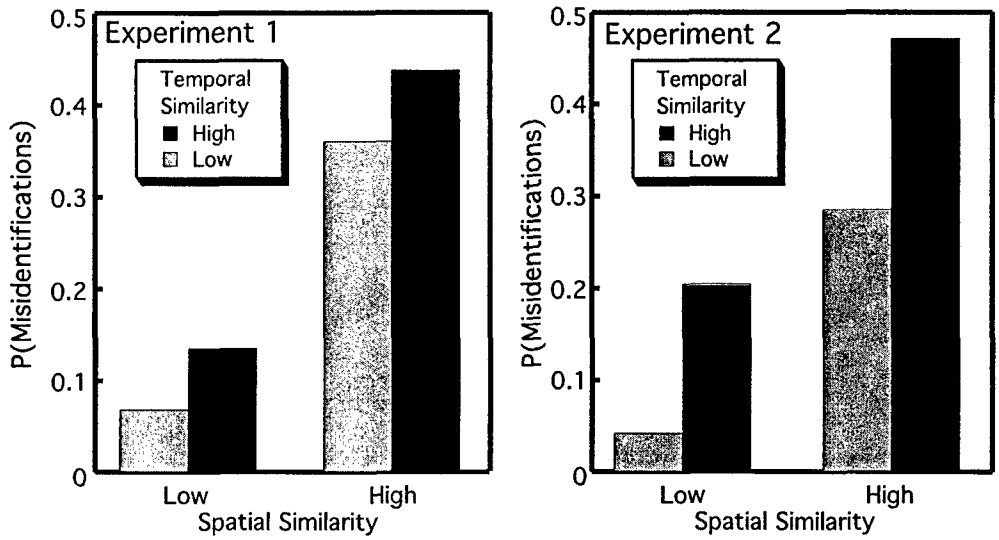

Figure 18. Proportion of misidentified items on T trials as a function of the spatial and temporal similarity between correct and misidentified study stimuli. Panel A. Misidentifications in Experiment 1. Panel B. Misidentifications in Experiment 2. Two levels of spatial difference are plotted on the x-axis. Each of them is plotted separately for stimulus pairs that were temporally similar (black bars), and for pairs that were temporally dissimilar (gray bars).

categories. In each panel, the horizontal-axis shows the two levels of spatial similarity; black bars show results with high temporal similarity, and gray bars show results with low temporal similarity. Note first that we can easily dismiss the hypothesis that all misidentifications resulted from a completely stochastic process. Monte Carlo simulation showed that a completely random process, which would produce correct identification of serial position on just 0.1875 of all trials, would generate values as extreme as the highest value in either panel of Figure 18 on fewer than one in 1,000,000 replications of the experiment. In fact, the distribution of misidentifications is consistent with the alternative hypothesis, namely that both temporal and physical similarity induced source errors, with physical similarity producing a larger effect than temporal similarity. Moreover, the two panels in Figure 18 shows no evidence of an interaction between the two variables, that is, the black and gray bars at low spatial similarity differ by about as much as the corresponding bars at high spatial similarity in both Experiments 1 and 2. Although the errors induced by perceptual similarity were larger than those induced by temporal similarity in both experiments, the power of temporal similarity seems stronger in Experiment 2. The one difference between Experiment 1 and 2, the proportion of T trials, could explain this difference in results.

The summed similarity of **p** and study items is an effective predictor of recognition memory of stimuli like those used here (Kahana & Sekuler, 2002; Nosofsky & Kantner, 2005). Figure 19 illustrates in schematic form some key elements of NEMo, the summed similarity model proposed by Kahana and Sekuler (2002). NEMo assumes that that study items, $\mathbf{s}_1$ $\mathbf{s}_2$, $\mathbf{s}_3$, are stored in memory as corresponding noisy exemplars, $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$, where exemplars' subscripts signify the order in which the stimuli were presented. NEMo computes pairwise similarities, η_1, η_2, η_3, between **p** and the noisy exemplar of each study item. If the sum, Σ_η, of these pairwise similarities exceeds an (optimal) criterion, the model responds that **p** had been in the study series. If Σ_η fails to exceed the criterion value, the model responds that **p** was not among the items in the study series. In NEMo's computation, sets of study and **p** items, together with random noise in the exemplar representations produce a distribution of values of Σ_η. To enhance the transparency of the following line of argument and to avoid having to estimate noise parameters for this purpose, we will put aside the effects of noise, treating exemplars as though they were noise free. In addition, for the sake of generality we will ignore β, a parameter that distinguishes NEMo from other summed similarity models; β captures the similarity of each study item to the others (Kahana & Sekuler, 2002; Nosofsky & Kantner, 2005). Neither of these simplifying departures from NEMo is consequential for what follows.

By definition, on T trials one study item matched **p**. As a result of this maximum similarity between one of the study items and **p**, Σ_η tends to be higher on T trials than on L trials. Note that this difference in Σ_ϕ for T and L trials will vary over trials, depending upon the particular study items and **p** presented on each trial. However, the mean difference in summed similarity between the two trial types will produce greater than chance success in recognition. The match between **p** and one of the study items has another consequence. On T trials one value of η_i will always be constant, at the upper limit of possible similarity values. On T trials only two values of η_i are free to vary, whereas on L trials all three values vary. Therefore, the variance in Σ_η, for T trials will tend to be smaller than the variance in Σ_η for L trials. In order to motivate the generation of ROC curves, we can cast these distributional relationships into signal detection terms, describing the distribution of Σ_η for T trials as the *signal* distribution, and the distribution of Σ_η for L trials as the *noise* distribution.

To examine the link between zROC slope and the distributions of Σ_η values on T and L trials, we calculated the summed similarity for each trial in our study. To minimize assumptions needed for the calculation, we substituted pairwise Euclidean spatial frequency differences between stimuli for the corresponding perceptual differences specified by NEMo. Although these two variables, physical and perceptual distance, are most likely related by an exponential transform (Kahana & Sekuler, 2002), any increasing monotonic relationship between the two variable leaves the argument unchanged. **p**-study summed distances were calculated separately for T trials and for L trials in Experiment 3. The frequency distributions of Σ_η for all stimulus sets that appeared in Experiment 3 are plotted in Figure 20A. Note that the x-axis

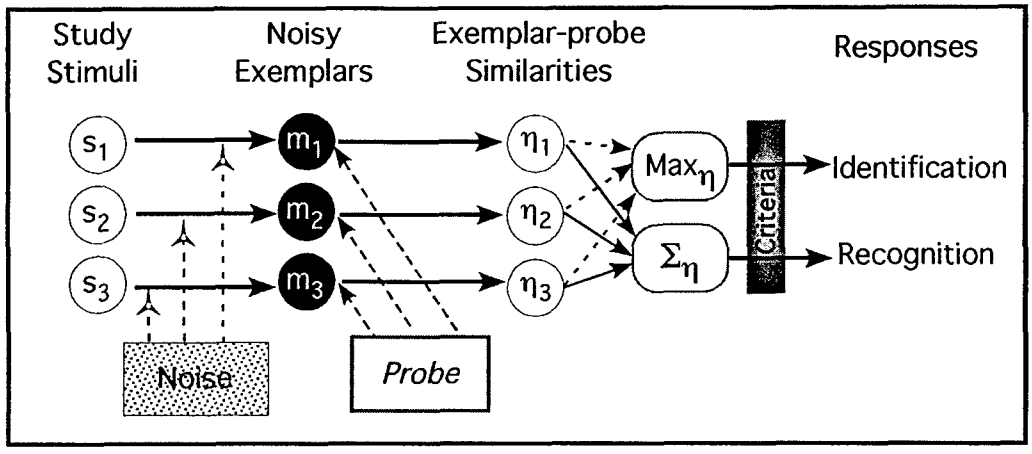

Figure 19. Schematic of model showing key stages that could lead to recognition and identification responses. Samples of noise are added to visual representations of the study stimuli ($\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$) producing a set of corresponding, noisy exemplars ($\mathbf{m}_1$, $\mathbf{m}_2$, $\mathbf{m}_3$) which are stored in memory. At the presentation of the probe stimulus, the similarity between $\mathbf{p}$ and each memory representation is computed and stored as η_1, η_2, η_3. These separate similarity measurements are combined into a summed similarity value, Σ_η. Omitted from this schematic representation are parameters that transform physical, stimulus distances into perceptual similarity, and β, which captures variation in the overall similarity of study items to one another. So that the model can also identify the serial position of the study item that matched $\mathbf{p}$, one might add a max operator that returns the serial position associated with the largest member of the similarity set, η_1, η_2, η_3.

has been reversed so that the smallest value of summed distance lies to the right. Because summed similarity, Σ_η, and summed distance are inversely related, the reversal of the normal x-axis direction represents increased summed similarity from left to right, and also brings the visual format of Figure 20A's distributions into conformity with formats commonly used in signal detection theory. As expected, the mean value of Σ_ϕ for T trials tended to be larger than the comparable value for L trials; also, the distribution of Σ_η for L trials had larger variance than did Σ_η for T trials. This unequal variance in the empirical signal and noise distributions is consistent the fact that recognition zROC slopes were > 1.0 (Wickens, 2002).

To demonstrate that distributional differences in Σ_η for T and L trials would actually have the predicted effect on zROC slope, we applied a signal detection analysis to the recognition data that would have been produced by several different distributions of Σ_η. We simulated recognition for several different subsets of stimuli, which were derived from the stimulus sets actually used in Experiment 3, the one experiment in which recognition was measured directly. First, we calculated the value of Σ_η for all 400 L trials and for all 400 T trials that each subject saw. Figure 20A shows

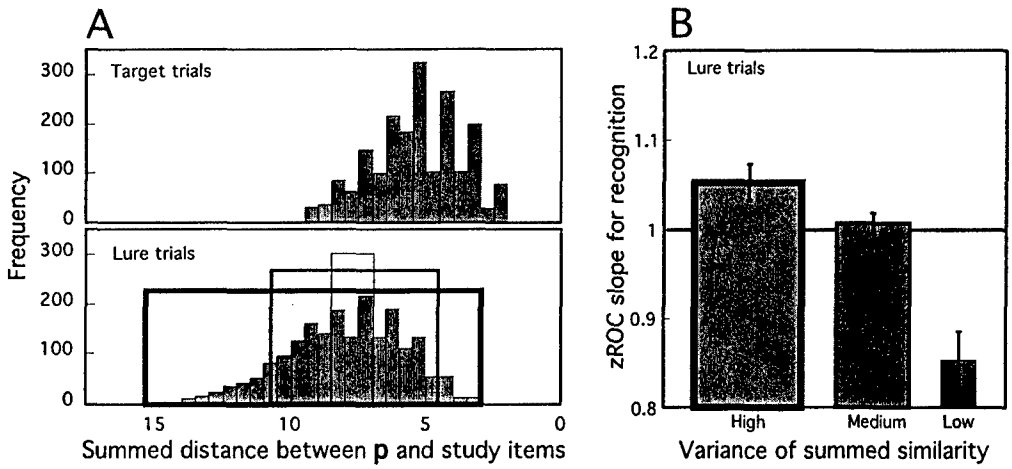

Figure 20. Panel A: Distributions of summed similarity values for all T trials (upper sub-panel) and all L trials (lower sub-panel) used in Experiment 3 and in simulations of recognition. L trials used both in Experiment 3 and in the first simulation are bracketed by the rectangle with the thickest line. The two narrower rectangles bracket L trials used in the second and third simulations. Panel B: Mean and SeM slopes of zROCs calculated for three sets of L trials; the variance of L trials' summed similarity values varied systematically among the three sets of trials. The rightmost point shows the mean and SeM zROC slope simulated for all stimulus sets used in Experiment 3. The thick, medium, and thin bars each correspond the simulations with large, medium, and small variance of summed similarity. The thickness of lines in Panel B corresponds to those in Panel A. The standard deviation of the T distribution is 1.68; the standard deviation for L trials was 2.1.

the distributions of Σ_η for the two trial types, aggregated over all subjects. Note that because stimulus sets were generated randomly from the pool of 25 items, each subject had been tested with a partially unique set of T and L stimuli. From each subject's responses to his or her 400 T and 400 L trials, we generated a zROC curve and calculated its slope. Note that these 800 trials were the actual stimulus set used in Experiment 3, the mean slope shown in the leftmost point in Figure 20B was 1.09, a value identical to the corresponding value obtained in the actual experiment (see rightmost point in Figure 17B).

Having confirmed that the variances of distributions of Σ_ϕ differed between the complete set of T and L trials, we constituted two subsets of stimuli. For the first subset we sought to reverse the relationship between the distributions' variances, while holding constant the difference between distributions' means; we reasoned that these new, modified distributions should produce zROC slopes below one. Starting with each subject's original stimulus set of 400 T and 400 L trials we carved out a reduced-variance L distribution by selecting 60 L trials whose Σ_ϕ values were near the mean value of Σ_ϕ. In addition, 60 T trials were randomly selected without regard

to their value of Σ_ϕ. This maneuver reduced the variance for noise trials, leaving the variance for signals trials unchanged, while also preserving the mean difference between the two distributions. Altering the relative variances of Σ_ϕ for T and L trials had the expected effect on zROC slope: the relatively narrower distribution of summed similarities values on L trials produced a simulated zROC mean slope of 0.85, a value considerably below 1.0 (the rightmost value shown in Figure 20B).

Finally, we repeated the simulation with a somewhat larger subset of values sampled from the original distribution of Σ_η from L trials; the aim was to generate a sample of stimuli in which the variance of Σ_η for L trials was approximately the same as that for T trials, which should produce zROC slopes near 1.0. We drew 250 L trials and 250 T trials from the original sets of 400 items. L trials were drawn without replacement from a region in vicinity of the mean for all L trials, but T trials for the subset were drawn at random from the entire distribution of 400 trials. The result was a ratio of variances between distributions that was intermediate to the ratios for the 60-trial sets and for the complete, 400-trial sets. Again, in constructing these set of stimuli, we held constant the mean distance between these two distributions of Σ_η. From each subject's own responses in Experiment 3 to each of these stimulus lists, we generated a zROC recognition curve for that subject. The mean and standard error of the zROC slopes are shown by the middle point in Figure 20B. Note that as expected, the mean zROC slope here was intermediate to the mean slopes from the other two simulated conditions, and was close to 1.0.

As mentioned earlier, the recognition zROC slopes produced in our experiments were all > 1, a result that is inconsistent with most previous reports (for example, Ratcliff et al., 1992). In concert with expectations from signal detection theory (Egan, 1975; Wickens, 2002), the simulations represented in Figure 20A and B show that our result arose from the differences in the variances of signal and noise distributions. As we have demonstrated, the summed similarity values for L trials used trials in Experiment 3 had larger variance than the values for T trials, a fact that signal detection theory predicts, and our simulations confirm, would produce zROC slopes >1.0.

It is difficult to specify how our results could be extended to more commonly used memory materials including words. For one thing, it is difficult to know what metric space is appropriate to describe such materials, which makes it hard to quantify their similarity relations. However, our simulations do imply that one should not expect any standard or canonical value of recognition zROC slope. Instead, they suggest that zROC slope will almost certainly depend upon variables such as the particular stimulus values that constitute stimulus lists, the range and distribution of pairwise similarities represented in the stimulus pool, the algorithm used to make up stimulus lists from that pool, and the number of items in the study list. For example, with factors constant, the larger the study list, the smaller the impact on T trial summed similarity of the one study that matches $\mathbf{p}$. This diminished impact will reduce the difference in the variance in Σ_η associated with T and L trials.

Our ROC results clearly deviate from their counterparts in memory results for verbal materials. As already discussed, we find zROC slopes of 1.1-1.3 for recognition and 0.8 to 0.9 for source identification, whereas studies using verbal materials report slopes of 0.8 to 1.0 for recognition and approximately 1.0 for source identification. For recognition zROC slopes, it is likely that in our experiments lures are much more similar to study items than would be the case with commonly used verbal stimuli. For source recognition results, our stimuli and task differ in many ways from stimuli and task used with verbal materials, which makes it impossible to know which might actually be responsible for the difference in results. For example, our task used short lists of study items, and defined source in terms of serial position that an item occupied; analogous studies with verbal material have used considerably longer lists of study items, and have mostly defined source in terms of the voice, male or female, in the which the words were heard.

Source of misattribution errors Sensory researchers have long understood that perceptual errors, such as illusions, can be valuable sources of insight into perception's normal operation (Eagleman, 2001). In the same way, from the very beginning of systematic research on memory (Ebbinghaus, 1885/1913), errors and failures have been extraordinarily useful in illuminating memory's normal operation. Here, we have focused on one particular kind of error –misattribution or source errors (Johnson et al., 1993; Johnson, 1997).

The results in Figure 18 suggest that spatial and temporal similarity both influence misidentifications of serial position, and that there separate effects are approximately additive. How, though, might these influences operate? To understand how spatial similarity might lead to source errors, consider a recent account of false alarms in recognition judgments. As described before, Kahana and Sekuler (2002)'s NEMo model assumes that three study items, $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$, are stored in memory as noisy exemplars. When the probe, $\mathbf{p}$, is presented, $\eta_1 \ldots \eta_3$, the set of similarities between $\mathbf{p}$ and each of the noisy exemplars is computed. Again, subscripts signify the order in which study stimuli were presented originally. NEMo describes each similarity value as exponentially decreasing function of the spatial difference between $\mathbf{p}$ and the corresponding values, $\mathbf{m}_1$, $\mathbf{m}_2$, $\mathbf{m}_3$. From the resulting similarity values, the summed similarity, Σ_η, is computed. In NEMo, a value of $\Sigma_\eta > k$, where k is an optimal criterion, constitutes evidence that at least one of the study items matched $\mathbf{p}$, which makes $\mathbf{p}$ seem familiar. Over trials, the probability of a recognition response (that is, a *yes* response) corresponds to the proportion of trials on which $\Sigma_\eta > k$.

To implement judgments of serial position, we propose that NEMo could be expanded to allow the model to perform an additional operation on the set of pairwise $\mathbf{p}$-study item similarities it was already computing. We are not proposing here a full, formal, quantitative account of this modification, and considerable new data would be required to explore any such account. However, we do think it worthwhile to sketch out the key elements of what we see as one plausible approach to serial

position judgments that is compatible with a general, summed similarity approach.

In an expanded version of NEMo, each trial's pairwise similarities would processed with a max operator, which returns the index of the largest item in η_1, η_2, η_3. In the absence of error, this index would be the serial position of the study item that most closely matches **p**. The subject bases the serial position identification upon that returned index. Because of noise associated with each exemplar, there will be trials on which the index returned by max will not correspond to the serial position whose study item physically matched **p**, but would correspond instead to the serial position of another study item. On such trials, the would model generate a source error, misidentifying the serial position of the matching item's similarity of **p**. The probability of such errors will be some monotonically decreasing function of **p**. In other words, study items that did not match **p** but were perceptually similar to it would be more likely misidentified as the match than would study items that were less similar to **p**. This is the pattern of results shown in Figure 18.

A different mechanism is required to motivate the other, temporal source of misidentifications in our results. Drawing upon an account of temporal effects in free recall (Howard & Kahana, 1999, 2002), we assume that the representation of each noisy exemplar is tagged in memory with a temporal code. We assume further that each item's temporal tag can be degraded as a result of passage of time and/or interference. If this degradation were partial rather than complete (Dodson, Holland, & Shimamura, 1998), serially adjacent positions in a sequence would more likely be confused with one another than would be positions more widely separated in a sequence. In the case at hand, with partial loss of serial position information, $\mathbf{s}_1$ is more likely misremembered as $\mathbf{s}_2$ than as $\mathbf{s}_3$, and $\mathbf{s}_3$ is more likely misremembered as $\mathbf{s}_2$ than as $\mathbf{s}_1$. This account is mute on errors involving forgetting of $\mathbf{s}_2$'s serial position. Again, this is the pattern of results seen in Figure 18. We should note that this effect resembles Dodson et al. (1998)'s demonstration that even when subjects misidentify the source of information, they sometimes retain partial information about that source.

We should note that the max operator with which NEMo might be expanded is kin to max operators demonstrated in several other contexts. For example, Gawne and Martin (2002) showed that some neurons in primate visual cortex behave in accordance with a max operator when two stimuli are present simultaneously in the receptive fields of such neurons. Closer to our own use of a max operator is the important role that such an operator plays in competitive queuing in planning and producing serially-ordered behaviors (Bullock & Rhodes, 2003; Rhodes, Bullock, Verwey, Averbeck, & Page, 2004; Agam, Bullock, & Sekuler, 2005). It is noteworthy that competitive queuing often leads to transposition errors, a counterpart to the temporal order transpositions demonstrated in our experiments. Although other, alternative processes may ultimately prove to be responsible for serial position identification, we believe that a max operator could quite plausibly participate in that process.

Across the three experiments reported here, zROC curves generated for recog-

nition performance had slopes >1. This was true in Experiments 1 and 2, where recognition performance was estimated from source identification judgments, as well as in Experiment 3, where recognition performance was measured directly, that is, from recognition judgments. Analyzing the distributional characteristics of T and L trials we demonstrated that zROC slopes were well accommodated by integrating a summed similarity computation for episodic recognition and a signal detection framework for decision making.

Personnel Supported by Project

- Robert Sekuler, Principal Investigator
- Michael J. Kahana, Consultant
- Yuko Yotsumoto, Graduate Research Assistant
- Feng Zhou, Graduate Research Assistant
- Chris McLaughlin, Research Assistant

Publications and Presentations

Results from this project have been published in Zhou et al. (2004); two other manuscripts have been submitted, one is under review, and the other is in press at the journal Memory & Cognition. One of these manuscripts reports the work described above, under Objective One; the other manuscript reports work done under AFOSR support builds on the paradigm introduced in (Zhou et al., 2004) and applies the modeling techniques developed by (Kahana & Sekuler, 2002). In addition to these published or submitted manuscripts, the PI and graduate researchers on the project have presented subsets of the results at meetings of the Visual Sciences Society in 2003, 2004 and 2005, the Mathematical Psychology Society in 2004, the European Conference on Visual Perception in 2003, and the Psychonomics Society in 2004. In addition, the PI took part in the AFOSR Cognition Program Review in 2004.

Interactions/Transitions

Publications relating to this project (Kahana & Sekuler, 2002; Zhou et al., 2004) and our many presentations at professional meetings have attracted considerable interest among vision researchers and among researchers interested in memory. As a result of learning about our work and discussions with the PI and his collaborators, Rob Nosofsky (Indiana University), a leading researcher in categorization and memory, has replicated key aspects of NEMo, our model for episodic recognition memory (Nosofsky & Kantner, 2005). In addition, following up the AFOSR Program Review, the PI launched a small scale effort to explore the online visuospatial memory paradigm presented at the Arizona Program Review by Don Lyon and Kevin Gluck (Air Force Research Laboratory). To facilitate this effort, Lyon generously gave the PI access to task software developed at AFRL; it is expected that this effort will continue even though AFOSR support has ended.

Inventions

. None.

References

Agam, Y., Bullock, D., & Sekuler, R. (2005). *Imitating unfamiliar sequences of connected linear motions* (Tech. Rep. No. 2005-3). Brandeis University, Waltham MA: Volen Center for Complex Systems.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Science, 5*(May), 204-210.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 443-446.

Brunswik, E., & Reiter, L. (1937). Eindruckscharaktere schematisierter gesichter. *Zeitschrift für Psychologie, 142*, 67-134.

Bullock, D., & Rhodes, B. (2003). Competitive queuing for serial planning and performance. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (p. 241-244). Cambridge, MA: MIT Press.

DeCarlo, L. (2002). Signal detection theory with finite mixture distributions: theoretical developments with applications to recognition memory. *Psychological Review, 109*(4), 710-721.

DeCarlo, L. (2003). An application of signal detection theory with finite mixture distributions to source discrimination. *Journal of Experimental Psychology: Learning, Memory & Cognition, 29*(5), 767-778.

Diamantaras, K. I., & Kung, S. Y. (1996). *Principal Component Neural Networks.* John Wiley & Sons, Inc.

Dobbins, I. G., Foley, H., Schacter, D. L., & Wagner, A. D. (2002). Executive control during episodic retrieval: multiple prefrontal processes subserve source memory. *Neuron, 356*(5), 989-996.

Dodson, C., Holland, P., & Shimamura, A. (1998). On the recollection of specific- and partial-source information. *Journal of Experimental Psychology: Learning, Memory & Cognition, 24*, 1121-1136.

Donaldson, W., & Murdock, B. B. (1968). Criterion change in continuous recognition memory. *Journal of Experimental Psychology, 76*, 325-330.

Druzgal, T. J., & D'Esposito, M. (2001). Activity in fusiform face area modulated as a function of working memory load. *Cognitive Brain Research, 10*(3), 355-364.

Druzgal, T. J., & D'Esposito, M. (2003). Dissecting contributions of prefrontal cortex and fusiform face area to face working memory. *Journal of Cognitive Neuroscience, 153*(6556), 771-784.

Dryden, I. L., & Mardia, V. (1998). *Statistical Shape Analysis.* New York: J Wiley.

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia, 41.*(6), 713-720.

Eagleman, D. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience, 2*(12), 920-926.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). Teachers College, Columbia Univerity.

Egan, J. P. (1975). *Signal Detection Theory and ROC-analysis.* New York: Academic Press.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the Face.* Oxford: Prentice-Hall.

Ennis, D. M. (1988). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General, 117*(4), 408-411.

Ennis, D. M., Mullen, K., Frijters, J. E. R., & Tindall, J. (1989). Decision conflicts: Within-trial resampling in Richardson's method of triads. *British Journal of Mathematical & Statistical Psychology, 42*, 265-269.

Garner, W. R., & Hake, H. W. (1951). The amount of information in absolute judgements. *Psychological Review, 58*(6), 446-459.

Gawne, T. J., & Martin, J. M. (2002). Responses of primate visual cortical v4 neurons to simultaneously presented stimuli. *Journal of Neurophysiology, 88*(3), 1128-35.

Geisler, W. S., & Chou, K. L. (1995). Separation of low-level and high-level factors in complex tasks: visual search. *Psychological Review, 102*(2), 356-378.

Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memoyr & Cognition, 30*(6), 1176-1195.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research, 39*, 3537-3560.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimentl Psychology: General, 123*(2), 178-200.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585-612.

Graham, N. V. S. (1989). *Visual Pattern Analyzers.* New York: Oxford University Press.

Hilford, A., Glanzer, M., Kim, K., & DeCarlo, L. (2002). Regularities of source recognition: ROC analysis. *Journal of Experimental Psychology: General, 131*(4), 494-510.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory mode. *Psychological Review, 95*, 528-551.

Hockley, W. E., & Cristi, C. (1996). Tests of the separate retrieval of item and associative information using a frequency-judgment task. *Memory & Cognition, 24*(6), 796-811.

Hole, G. J. (1996). Decay and interference effects in visuospatial short-term memory. *Perception, 25*(1), 53-64.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learnng, Memory & Cognition, 25*(4), 923-941.

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language, 46*(1), 85-98.

Hwang-Grodzins, G., Jacobs, J., Danker, J., Sekuler, R., & Kahana, M. J. (2005). Neural correlates of sub-vocal rehearsal in a modified Sternberg task. In *Proceedings of the Annual Meeting of the Cognitive Neuroscience Society.* New York.

Johnson, M. (1997). Source monitoring and memory distortion. *Philosophical Transactions of the Royal Society, London Series B, 352*(1362), 1733-1745.

Johnson, M., Hashtroudi, S., & Lindsay, D. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3-28.

Johnstone, A. B., & Williams, C. (1997). Do distinctive faces come from outer space? an investigation of the status of a multidimensional face-space. *Visual Cognition, 4*(1), 59-67.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research, 42*, 2177-2192.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302-4311.

Kerr, J., Ward, G., & Avons, S. E. (1999). Response bias in visual serial order memory. *Journal of experimental Psychology: Learning, Memory & Cognition, 24*(5), 1316-1323.

Klein, I., Paradis, A. L., Poline, J. B., Kosslyn, S. M., & Le Bihan, D. (2000). Transient activity in the human calcarine cortex during visual-mental imagery: an event-related fmri study. *Journal of Cognitive Neuroscience, 12*(Supplement 2), 15-23.

Kosslyn, S., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature, 378*(6556), 496-498.

Lee, C., & Estes, W. (1977). Order and position in. primary memory for letter strings. *Journal of Vebal Learning & Verbal Behavior, 16*, 395-418.

Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space hypothesis. *Psychological Science, 11*(5), 379-385.

Lehky, S. R. (2000). Fine discrimination of faces can be performed rapidly. *Journal of Cognitive Neuroscience, 12*(5), 848-855.

Levin, D., & Beale, J. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception & Psychophysics, 62*(2), 386-401.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within subject designs. *Psychonomic Bulletin & Review, 1*, 476-490.

Magnussen, S., & Greenlee, M. W. (1999). The psychophysics of perceptual memory. *Psychological Research, 62*(2-3), 81-92.

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Eaperiment Psychology: Human Perception & Performance, 22*, 294-317.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms.* Cambridge: MIT Press.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*(89), 609-626.

Nachmias, J., & Steinman, R. M. (1963). Study of absolute visual detection by the rating scale-method. *Journal of the Optical Society of America, 53*, 1206-1206.

Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research, 39*(23), 3824-3833.

Nosofsky, R. M. (1986). Attention, similarity, and the identification categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance, 17*(1), 3-27.

Nosofsky, R. M., & Kantner, J. (2005). Familarity, study-list homogeneity, and short-term perceptual recognition. *Memory & Cognition, in press.*

Olzak, L. A., & Thomas, J. P. (1999). Neural recoding in human pattern vision: model and mechanisms. *Vision Research, 39*(23), 231-256.

Page, M. P., & Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychological Review, 105*, 761-81.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.

Pelli, D. G., Robson, J. G., & Wilkins, A. J. (1988). Designing a new letter chart for measuring contrast sensitivity. *Clinical Vision Sciences, 2*, 187-199.

Peters, R. J., Gabbiani, F., & Koch, C. (2003). Human visual object categorization can be described by models with low memory capacity. *Vision Research, 43*(21), 2265-2280.

Phillips, W., & Christie, D. (1977). Components of visual memory. *Quarterly Journal of Experimental Psychology, 29*, 117-133.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics, 16*, 283-290.

Phillips, W. A. (1983). Short-term visual memory. *Philosophical Transactions of the Royal Society B, 302*, 295-309.

Principle, J. C., Euliano, N. R., & Lefebvre, W. C. (2000). *Neural and Adaptive Systems.* John Wiley & Sons, Inc.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518-535.

Rhodes, B. J., Bullock, D., Verwey, W. B., Averbeck, B. B., & Page, M. P. A. (2004). Learning and production of movement sequences: Behavioral, neurophysiological, and modeling perspectives. *Human Movement Science, 23*, 699-746.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science, 4*, 28-34.

Rotello, C., Macmillan, N., & Reeder, J. (2004). Sum-difference theory of remembering and knowing: a two-dimensional signal-detection model. *Psychological Review, 111*(3), 588-616.

Rugg, M. D., & Yonelinas, A. (2003). Human recognition memory: a cognitive neuroscience perspective. *Trends in Cognitive Science, 7*(7), 313-319.

Sekuler, R., Kahana, M. J., McLaughlin, C., Golomb, J., & Wingfield, A. (2005). Preservation of episodic visual memory in aging. *Experimental Aging Research, 31*, 1-12.

Sigala, N., Gabbiani, F., & Logothetis, N. K. (2002). Visual categorization and object representation in monkeys and humans. *Journal of Cognitive Neuroscience, 14*(2), 187-98.

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 26*(6), 1499-1517.

Sternberg, S. (1966). High speed scanning in human memory. *Science, 153*, 652-654.

Sternberg, S. (1975). Memory scanning: New findings and current controversies. *Quarterly Journal of Experimental Psychology, 27*, 1-32.

Torgerson, W. S. (1958). *Theory and Methods of Scaling.* J. Wiley & Sons.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of experimental Psychology, A, 43*(2), 161-204.

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception, 15*(5), 525-535.

Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: the effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology, A, 44*(4), 671-703.

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition, 30*(6), 885-892.

Watson, C. S., Rilling, M. E., & Bourbon, W. T. (1964). Receiver-operating characteristics determined by a mechanical analog to the rating scale. *Journal of the Acoustical Society of America, 36*, 283-288.

Weller, S. C., & Romney, A. K. (1988). *Systematic Data Collection* (Vol. 10). Newbury Park, CA: Sage Publications.

Wexler, K. N., & Romney, A. K. (1972). Individual variations in cognitive structures. In A. K. Romney, R. M. Shepard, & M. Nerlove (Eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences* (2 ed., p. 73-92). Seminar Press.

Wickens, T. D. (2002). *Elementary Signal Detection Theory.* New York: Oxford University Press.

Wilson, H. R., Loffler, G., & Wilkinson, F. (2002). Synthetic faces, face cubes, and the geometry of face space. *Vision Research, 42*, 2909-2923.

Yonelinas, A. P., & Levy, B. J. (2002). Dissociating familiarity from recollection in human recognition memory: different rates of forgetting over short retention intervals. *Psychonomic Bulletin & Review, 9*(3), 575-582.

Yotsumoto, Y., & Sekuler, R. (in pres). Out of mind, but not out of sight: Intentional control of visual memory. *Memory & Cognition.*

Zhou, F., Kahana, M. J., & Sekuler, R. (2004). Short-term episodic memory for visual textures: A roving probe gathers some memory. *Psychological Science, 153*(2), 112-118.